Molnár Gyöngyvér
Szegedi Tudományegyetem, Pedagógia Tanszék,
MTA-SZTE Képességkutató Csoport

A Rasch-modell alkalmazása a társadalomtudományi kutatásokban

A teljesítményt mérő skálák megalkotásának nehézsége abban rejlik, hogy azok látens, azaz rejtetten, fizikailag nem mérhető, nem látható tulajdonságok leírását célozzák meg, amelyek nem egyetlen egy változó függvényei, hanem bonyolult, összetett változórendszerrel leírható tulajdonságok.

Mindenkinek ismerősek a fizikai világ mérési skálái, akár a tegnapi minimum hőmérsékletről, akár egy gyerek magasságáról vagy életkoráról van szó. Mindegyik esetben megadunk egy mennyiséget, ami egy bizonyos skálán helyezhető el. Például ha a tegnapi minimum hőmérséklet Szegeden 10 fok volt, mindenki el tudja dönteni, hogy hideg vagy meleg volt Szegeden anélkül, hogy el kellene utaznia Szegedre az idő megvizsgálásához. Ha valaki 1 km-re lakik a buszpályaudvartól, mindenkinek van egy elképzelése arról, hogy az illető milyen messze lakik az érintett helytől anélkül, hogy legyalogolná az adott távolságot. Fizikai világunk tele van különféle skálákkal, amelyek hasznos információkkal látnak el mindenkit a minket körülvevő világról.

A természettudományok területén kívül is találkozhatunk skálákkal, bár azok nem annyira egyetemesek (például amikor az orvos a depresszió különböző szintjeiről beszél – ez jelentéssel bírhat a többi doktor, illetve az érintett paciens számára, de aki nem jártas a témában, nem érti.) A tanár osztályzatokat ad a diákjainak, ami az adott diák iskolai előrehaladását mutatja, de általánosságban a normaorientált, szubjektív osztályozás miatt nem lehet messzemenő következtetéseket levonni a jegy értékéből. Ahhoz, hogy egy univerzális skálát megalkossunk az osztályozások területén, tudni kellene, hogy pontosan mit tud a diák, és definiálnunk kellene az optimális iskolai teljesítményt. Előbbit különböző tesztekkel próbáljuk becsülni, azonban a teljesítményt ezáltal mindig bizonyos változók, feladatokon nyújtott teljesítmény alapján állapítjuk meg, utóbbit pedig – ha iskolában tanultakról van szó – a NAT és kerettantervi szabályozásokon keresztül próbáljuk megközelíteni. A skálák megalkotásának további nehézsége, hogy a mérések adatai különböző skálákon helyezkednek el (nominális, ordinális, intervallum és arány), amit az eredmények értelmezése során szem előtt kell tartani. Ha az adatok arányskálán helyezkednek el, akkor beszélhetünk az eredmények közötti különbségek nagyságáról és arányáról is, ha az adatok intervallumskálán vannak, akkor már csak a számok közötti különbségek nagyságát értelmezhetjük. Ordinális skála esetén csak rangsorról, és nem távolságokról, nominális skála esetén pedig csak az adott érték nominális értékéről beszélhetünk. Ha egymáshoz viszonyítanánk ezeket a skálákat, akkor a legalacsonyabb szinten a nominális skála, felette az ordinális skála lenne, azt pedig az intervallum és az arány skála követné. Ahogy egyre magasabb szintre érünk, egyre bővül az elvégezhető matematikai műveletek köre. Ebből adódóan, ha látens struktúrákat jellemző skálákat szeretnénk fejleszteni, a legjobb, ha a legmagasabb skálatípust, azaz az arányskálát alkalmazzuk. Ennek egyik nehézsége, hogy nehéz meghatározni az abszolút nulla pontot, azaz azt, hogy mikor nem beszélhetünk az adott látens tulajdonság létezéséről. Ebből adódóan egy rejtett tulajdonság skálájának megalkotásakor elegendő, ha arra törekedünk, hogy adataink intervallumskálán legyenek.

Ideális mérés esetén, ha kiválasztunk három diákot, például Annát, Bélát és Csabát, és az egyik diák, Anna a teszten kevés pontot ért el, Béla Annánál néhány ponttal többet és Csaba magas pontszámot, akkor egy másik azonos képességet mérő teszten is hasonlónak kell lennie az elért pontok eloszlásának. Azaz Anna keveset, Béla kicsivel többet, Csaba pedig sokkal több pontot érne el. Ha fennállna ez az eset, akkor az adott képesség ezen eredmények alapján megalkotott képességskálája intervallumskála lenne, hiszen nemcsak a diákok sorrendjéről, hanem képességszintbeli távolságukról is beszélhetnénk.

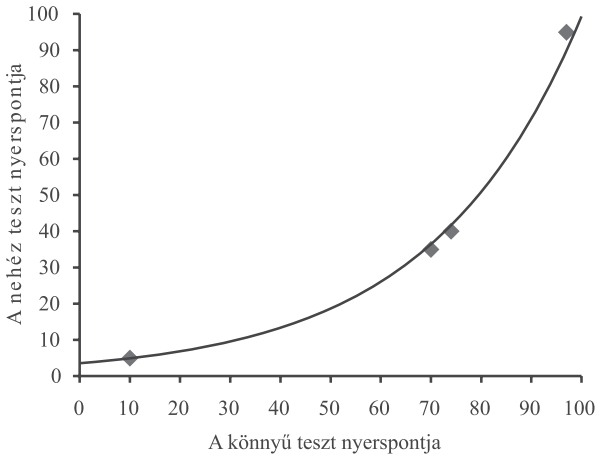
Képességszintek meghatározása

A képességszintek meghatározása a nyerspontokból

A klasszikus tesztelmélet eszközrendszerével minden esetben vagy a diákok nyerspontjait, vagy azokat százalékos formában kifejezve hasonlítjuk egymással össze. A következőkben egy modellált mérés segítségével bemutatom ezen összehasonlítási módok hátrányát.

Tegyük fel, hogy egy képesség mérésére rendelkezésünkre áll két teszt, egy könnyebb és egy nehezebb. Az egyszerűség kedvéért tegyük fel, hogy az elérhető maximum pont mindkét teszten 100 pont, továbbá tegyük fel, hogy A, B, C és D négy különböző képességű diák a modellált populációból. A nagyon tehetséges, B nagyon alacsony képességszintű, C és D átlagos képességű diák. A továbbiakban megnézzük, hogyan alakul a négy diák egymáshoz viszonyított teljesítménye nyerspontjaik, illetve az IRT szemszögéből, ha a könnyű tesztet oldják meg, és milyen pontszámot érnének el, ha a nehéz tesztet oldanák meg.

A magas képességszintű diák (A) valószínű mind a könnyű, mind a nehéz teszten jól teljesít, magas pontszámot ér el, mivel képességszintje magasabb, mint a tesztek feladatainak megoldásához szükséges képességszint. Az alacsony képességszintű diák (B) valószínű mindkét teszten rosszul teljesít, mivel képességszintje alacsonyabb, mint a feladatok átlagos megoldásához szükséges képességszint. Az átlagos képességszintű diákok (C és D) teljesítményét azonban erőteljesebben befolyásolja a teszt nehézsége. Egy könnyebb teszt esetén C és D diák relatíve magasabb pontszámot ér el, míg egy nehezebb teszt megoldása során relatív alacsonyabbat. Az *1. ábra* mutatja az A, B, C és D diák teljesítményének alakulását a teszt nehézségének függvényében. A vízszintes tengelyről, ami a könnyű teszt pontszámait mutatja, leolvasható, hogy a könnyű teszten mutatott teljesítmény alapján A és C diák teljesítménye (nyerspont-értékben) közelebb áll egymáshoz, mint a B és D diáké, holott a függőleges tengelyen, ami a nehéz teszt nyerspontjait ábrázolja, a B és D diák teljesítménye közelebb áll egymáshoz, mint az A és C diáké. Ha viszont mindkét teszt ugyanazt a képességet méri, akkor elvárjuk, hogy mindkét skálán ugyanaz a képességszintbeli távolság legyen az A és C diák között. Ebből adódóan a tesztek nyerspont-értékei csak a diákok egymáshoz viszonyított sorrendjéről ad információt, de a közöttük lévő képességszintbeli távolságról nem.



1. ábra. Egy modellált populáció könnyű és nehéz teszten mutatott teljesítménye (Wu, 2006a alapján)

Technikai értelemben ebből következőleg azt mondhatnánk, hogy a nyerspontérté-

kek ordinális skálán helyezkednek el és nem intervallumskálán, azonban az sem teljesen fedné le a valóságot, mivel mindkét skálán közös, hogy a C és D diák képességszintje közelebb áll egymáshoz, mint a B és C diáké, azaz a valós skála, ha lenne ilyen skálafokozat, akkor valahol az ordinális és intervallumskála között lenne.

Egy másik fontos tény, hogy a két teszten elért pontszámok közötti kapcsolat nem lineáris, azaz a két skála egymásba transzformálásához nem elegendő egy lineáris transzformáció.

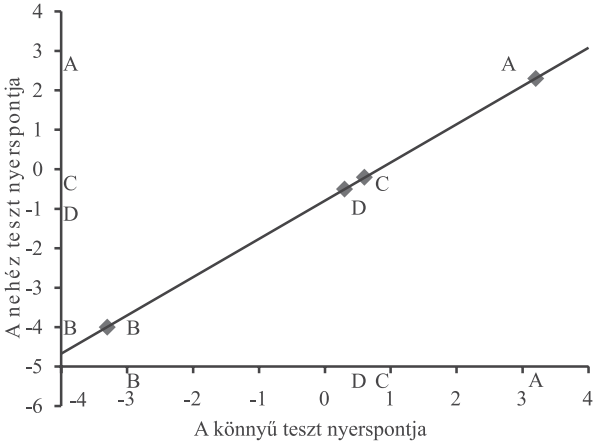
Hogyan változik a diákok egymáshoz viszonyított állása, ha nem a nyerspont-értékekkel, hanem azok százalékosan kifejezett formájával számolunk? Ez sem oldja meg a problémát, bár két különböző összpontszámú teszt eredményét össze tudjuk hasonlítani egymással, de a diákok között meglévő eredeti távolságok az eredmények százalékos formába való kifejezése során elvesznek.

A nyerspontok transzformációja az IRT szemszögéből

Az IRT (Item Response Theory) (Horváth György (1997) terminológiájával élve modern tesztelmélet) nem ekvivalens a Rasch modellel (*Rasch*, 1960), hanem a valószínűségi tesztelméletek egy gyűjtőfogalma, ahova a Rasch modellen és továbbfejlesztett változatain kívül még számos, más tesztmodell is besorolható. Ezek közül talán a legismertebb – bizonyos tulajdonságai miatt, amiket l. később – a Rasch-modell. Néhány más modellről magyarul részletesebben l. Horváth (1997) könyvét.

A valószínűségi tesztmodellek egyrészt abban különböznek egymástól, hogy milyen típusú összefüggést feltételeznek a személy képességparamétere és a helyes válasz valószínűsége között (pl: logisztikus függvény, normális eloszlásfüggvény), másrészt abban, hogy hány paraméterrel számolnak (l. pl. *Write és Masters,* 1982; *Molnár,* 2003). Mindegyik IRT modellben közös, hogy adott item esetén megadják a személy helyes válaszadásának valószínűségét, nem determinisztikusak, hanem valószínűségi alapokon nyugszanak, illetve ha ismert az itemek nehézségi indexe és a diákok képességparamétere, akkor megadják, hogy minden egyes diák milyen valószínűséggel oldaná meg jól külön-külön az egyes itemeket. A Rasch-modell alapvető elképzeléséről, matematikai hátteréről és tulajdonságairól l. a tanulmány későbbi alfejezetét.

A nyerspontok transzformációja során az a cél, hogy egy olyan matematikai függvényt találjunk, ami megszűnteti a teszt nehézségétől függő képességeloszlást. Erre egy alkalmas matematikai összefüggés az IRT modellek között egyedül, a Rasch-modellben használt logisztikus függvény. A Rasch-modell transzformációja a nyers adatokat egy olyan skálára transzformálja, ami már megőrzi a diákok közötti távolságok nagyságát is, azaz az *1. ábrá*n látható görbét ez a transzformáció kiegyenesíti. *(2. ábra)* A 2. ábrá n mind a vízszintes (könnyű teszten elért eredmény alapján meghatározott képességszint), mind a függőleges tengelyen (a nehéz teszten elért eredmény alapján meghatározott képes-



2. ábra. Egy modellált populáció könnyű és nehéz teszten mutatott képességszintjének alakulása

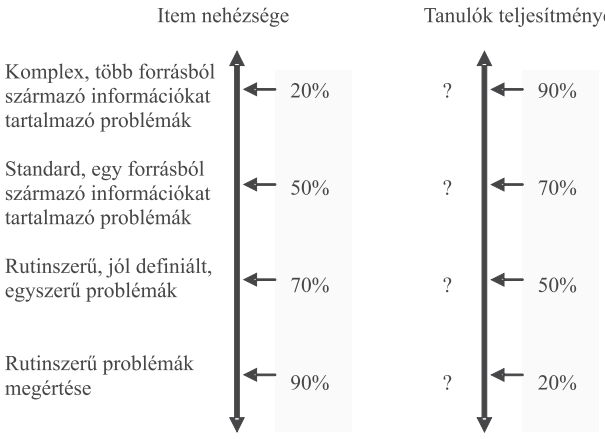
ségszint) az A és a C diák képességszintbeli távolsága megegyezik. (A transzformáció megőrzi a nyerspontok alapján kialakított sorrendet, ezért ha valakit csak a diákok sorrendje érdekel, nem kell IRT-hez folyamodnia.)

További problémaként merül fel, hogy a két skálán a képességszintek abszolút értékben különböznek egymástól. Ennek okáról l. később A Rasch-modell fő tulajdonságai alfejezetet.

Az itemek nehézségi szintjeinek és a diákok képességszintjeinek összekapcsolása

Egy ideális mérés során elvárjuk, hogy ha egy diák például 5 pontot ér el 100 pontból, akkor meg tudjuk mondani, hogy mit tud, az adott képesség fejlődésének milyen stádiumában van, mi várható el tőle, azonban ha nyers adatokat használunk a tanulók képességszintjének és az itemek nehézségi szintjének meghatározásakor, nem egyértelmű, hogy hogyan kapcsoljuk össze a két skálát.

Például egy problémamegoldó teszt itemnehézségi skáláján átlagosan a diákok 20 százaléka oldotta meg jól a bonyolultabb, komplexebb problémákat, amelyek megoldásához szükséges információk többféle forrásból származtak, míg a diákok 90 százalékának nem jelentett problémát a rutinszerű problémák megértése. *(3. ábra)* A diákok teljesítményét százalékban kifejező skálán is megvannak azok a pontok, ahol azon diákok állnak, akik átlagos teljesítménye 20, 50, 70, illetve 90 százalék egész teszten. A két skála százalékban megadott pontjai nehezen kapcsolhatóak össze.



3. ábra. Az itemek és személyek nehézség-, illetve képességszintjének összekapcsolása a nyerspontok alapján

Például igaz-e az, hogy aki 70 százalékos teljesítményt mutatott a teszten, akkor megoldja a standard, egy forrásból származó információkat tartalmazó problémákat. Ha nem ismert, hogy a teszt milyen arányban tartalmazza a fent nevezett problématípusokat, az is előfordulhat, hogy a teszt 70 százaléka rutinszerű problémák megoldásából áll, és ebben az esetben a 70 százalékos eredmény nem utal a jelen modell első lépcsőfokánál magasabb problémamegoldó képességbeli fejlettségi szintre. Ha ismerjük a teszt felépítését, akkor is nehéz diákokra lebontva mindenkihez hozzárendelni, hogy ki mit oldott meg, majd ennek alapján meghozni a döntést.

A képességek becslése az IRT eszközrendszerével

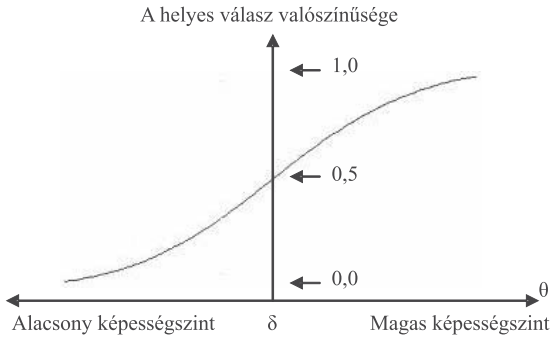
Amint korábban is utaltam rá, az IRT modellek egy-egy matematikai modellt használnak, ami a személy egy adott itemre adott helyes válaszának valószínűségét becsli figyelembe véve a személy képességparaméterét és az item nehézségét. Egy itemre adott helyes válaszok valószínűségét különböző képességszintek mellett az item item karakterisztikus görbéje írja le. *(4. ábra)*

A jó képességű diák jó válaszának valószínűsége közel áll 1-hez, míg az alacsony képességszintű diáké 0-hoz. Az átlagos diák – a modell értelmében – p=0,5 valószínűség-

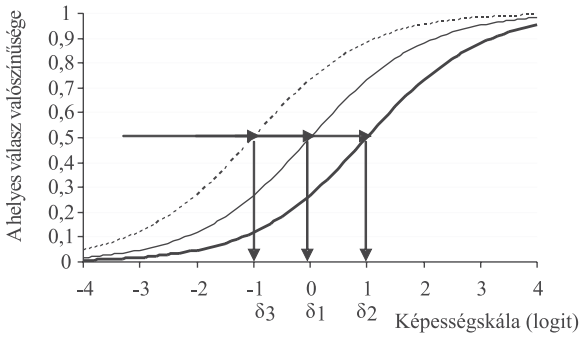
gel ad jó választ az itemre. A görbe megmutatja, hogy az egyes képességszintű diákok milyen valószínűség mellett válaszolnak jól az adott iremre (teszt-karakterisztikus görbe esetén, milyen képességszint mellett, hány pontot érnek el a teszten). A Rasch modellben az item nehézségét az adja meg, hogy milyen képességszint szükséges ahhoz, hogy p=0,5 legyen a helyes megoldás valószínűsége. Ez alapján a *4. ábrá*n az átlagos diák képességszintje (δ) adja az adott item nehézségi indexét. Ebben az értelemben az item nehézsége kapcsolatban áll a feladat nehézségével. A bemutatott item egy jó képességű diáknak könnyű, egy alacsony képességű diáknak nehéz, de az item nehézségét annak a diáknak a képességszintje határozza meg, aki 50-50 százalék valószínűséggel ront, illetve jól teljesít az itemen.

Az 5. ábra három különböző nehézségű item itemkarakterisztikus görbéjét ábrázolja. A három item nehézségi indexe: δ_1, δ_2, és δ_3. Jelen esetben δ_1=0 logit, δ_2=1 logit, és δ_3= -1 logit. Ha a görbék inflexiós pontjától (ahol a görbe gyorsuló növekedése lassulóvá vált át) húzunk egy-egy merőlegest az ordináta és abszcissza tengelyre, akkor leolvasható, hogy mindhárom görbe inflexiós pontjának ordináta koordinátája 0,5, azaz ha megnézzük a görbék inflexiós pontjában az abszcissza koordináta értékeket, leolvasható, hogy milyen képességszintű diák oldja meg 50 százalék valószínűséggel jól az adott itemet. A δ_1=0 nehézségű, azaz a középső itemkarakterisztikus görbe egy átlagos nehézségű görbe karakterisztikus görbéje, a δ_2=1, azaz a vastagított vonalú görbe egy átlagosnál 1 logitegységgel nehezebb item karakterisztikus görbéje, a δ_3= -1, azaz a szaggatott vonalú görbe, az átlagosnál 1 logitegységgel könnyebb item karakterisztikus görbéjét mutatja.

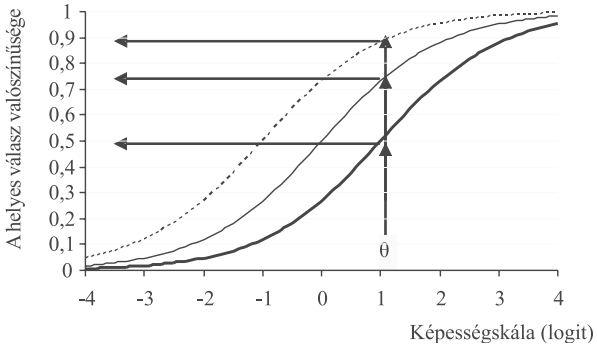
Miután az itemek nehézségi indexei a diákok képességszintjei alapján definiáltak, ezért az itemek nehézségét és a diákok képességszintjét közös képességskálán tudjuk ábrázolni. Ha ismerjük egy diák képességszintjét, meg tudjuk mondani, hogy milyen valószínűséggel oldana meg olyan itemet, amely nehéz-



4. ábra. Egy példa az itemkarakterisztikus görbére



5. ábra. Három különböző nehézségű item karakterisztikus görbéje



6. ábra. Egy θ képességszintű diák három különböző nehézségű itemre adott helyes válaszának valószínűsége

ségi indexe értelmezhető a közös képességskálán, anélkül, hogy a diáknak a valóságban meg kellene oldani azt. A *6. ábrá*n bemutatunk erre egy példát, ahol egy (θ) képességparaméterű diák három különböző nehézségű itemen való sikeres teljesítményének valószínűsége olvasható le.

Tegyük fel, hogy θ=1, ekkor a diák a δ=1 nehézségű itemet (vastagított vonalú itemkarakterisztikus görbéhez tartozó itemet) 50 százalék valószínűséggel oldja meg jól. Ugyanez a diák a δ=0 nehézségű itemet már 73 százalék valószínűséggel oldja meg, azaz közel 25 százalékal nagyobb valószínűséggel, míg a δ= -1 nehézségű itemet pedig 88 százalék valószínűséggel oldja meg jól. Az 1. táblázatban néhány logitban adott képességszint és nehézségi index mellett összefoglaltuk a helyes válasz valószínűségét és az adott esetből nyert relatív információ nagyságát.

A táblázatból leolvasható, hogy ha i-edik személy képességparamétere alacsonyabb, mint j-edik item nehézségi indexe, akkor a képességparaméter és a nehézségi index különbsége pozitív és a helyes válasz valószínűsége nagyobb, mint 50 százalék. Minél nagyobb ez a különbség, annál közelebb van a helyes válasz valószínűségének nagysága az 1-hez, azaz a 100 százalékhoz (a modell valószínűségi természetéből fakadóan, azt sohasem éri el). Ha az item túl nehéz az adott személy számára, azaz a képességparaméter és az itemnehézség különbsége negatív szám, akkor a sikeres megoldás valószínűsége kevesebb mint 50 százalék. Abszolút értékben minél nagyobb ez a különbség, annál közelebb lesz a helyes válasz valószínűsége 0-hoz.

1. táblázat. A helyes válasz valószínűsége a képességparaméter és az itemnehézségi mutató logitban adott függvényében (néhány példa)

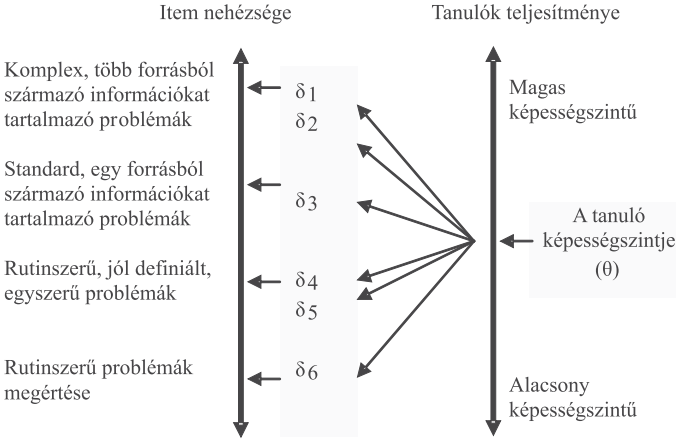
Képesség-paraméter θ_i	*Item-nehézség* δ_j	*Különbség* $(\theta_i-\delta_j)$	*A helyes válasz valószínűsége* p_{ij}	*Az adott információ erőssége (%)* $400*I_{ij}$	*Ugyanazon pontosság eléréséhez szükséges itemek száma*
5	0	5	0,99	4	250
4	0	4	0,98	7	142
3	0	3	0,95	19	53
2	0	2	0,88	45	23
1	0	1	0,73	79	13
0	0	0	0,50	100	10
0	1	-1	0,27	79	13
0	2	-2	0,12	45	23
0	3	-3	0,05	19	53
0	4	-4	0,02	7	142
0	5	-5	0,01	4	250

A táblázat utolsó előtti oszlopa arról ad információt, hogy az adott személy- és itemparaméterek mellett milyen mennyiségű relatív információt kapunk a személyről és itemről ($I_{ij} = p_{ij} (1-p_{ij})$). (*Write és Stone*, 1979) A kinyert relatív információt akkor tekintjük 100%-nak, ha a személy képességparamétere és az item nehézségi indexe megegyezik, azaz $\theta_i-\delta_j=0$. (*Write és Masters,* 1982) Ez a mérőszám annak eldöntését segíti, hogy az adott mérésben az adott item milyen mértékben járul hozzá a személy képességparaméterének meghatározásához, azaz mennyi információt veszítenénk az adott személy képességparaméterének meghatározása során, ha az adott itemet elhagynánk a tesztből. Ha az item nehézsége (δ_j) a személy képességparaméterének (θ_i) egy logitegységes környezetében van, akkor a δ_j és a θ_i –ról nyert információ több, mint 79 százalék, ami fokozatosan 45 százalékra csökken, ahogy a két paraméter közötti távolság nagysága 2 logitegységre nő. Ha ez a távolság több mint 3 logitegység, akkor az item hatásfoka kevesebb

mint 19 százalék, 4 logitegységnél nagyobb távolság mellett pedig már 7 százaléknál is kevesebb információt nyerünk az adott itemmel a személy képességparaméterét illetőleg (vagy fordítva, a személlyel, az item nehézségi indexét illetően). (Egy teszt információs függvényének meghatározásakor az egyes itemek információs függvényei összeadódnak ($I_t = \sum_{i,j=1}^{n} I_{ij}$, ahol I_t a teszt információs függvénye). (*Adema és Gademann,* 1992)

A táblázat utolsó oszlopa mutatja, hogy az aktuális képességparaméter és nehézségi index távolságában hány item szükséges a paraméterértékek ugyanazon pontossággal való meghatározásához. Minél nagyobb a személy képességparamétere és az item nehézségi indexe közötti különbség nagysága, annál több item szükséges a személy képességparaméterének minél pontosabb meghatározásához. Például 20 százalékos információt adó itemekből (ha θ_i-δ_j=1,8) öt darab szükséges ugyanazon pontosság eléréséhez, amit egy 100 százalékos itemmel érnénk el. Ha a két paraméter 3 logitegységes távolságban van egymástól, akkor 4-5-ször annyi itemet tartalmazó tesztre lenne szükség, mint ha a teszt itemeinek nehézségi indexe a személyparaméterek 1 logitegységes távolságán belül lennének.

Az itemnehézségi mutatók és a képességszintek közös skálán való ábrázolása lehetőséget ad arra, hogy a 3. ábra két skáláját össze tudjuk olvasztani. Ennek sematikus képét ábrázolja a *7. ábra*. (Empirikus adatokkal történő elemzését lásd például *Molnár,* 2004) Az itemnehézségi skálát (bal oldal) és a képességskálát (jobb oldal) a sikeres válaszadás valószínűségének matematikai függvénye kapcsolja össze. A θ képességszintű tanulóhoz minden egyes item esetén hozzá lehet rendelni egy valószínűséget, amilyen valószínűség mellett ő sikeresen oldja meg az adott itemet. Ennek következtében minden egyes diákhoz hozzárendelhető annyi valószínűségi szint, ahány itemről van szó, illetve minden egyes itemhez hozzárendelhető annyi diák képességszintje, akik a minta részét képezik. Például egy 20 itemből álló teszt esetében, amit 25 diák old meg, 500 (20x25) diák-item találkozást regisztrálhatunk, amelyekhez minden esetben az IRT modellek kiszámolják a helyes válaszadás valószínűségi szintjét, majd ezeket a valószínűségeket használják fel minden egyes diák elvárt teljesítményének és válaszmintázatának meghatározásakor, illetve minden egyes item minden tanulóhoz való hozzárendelése során is. (*Griffin*, 1999)



7. ábra. Az itemek és személyek összekapcsolása az IRT segítségével

Ha minden egyes item esetén le tudjuk írni, hogy milyen képességek szükségesek megoldásukhoz, akkor könnyen meg tudjuk határozni, hogy egy adott képességszintű diák milyen szinten van az adott képességterületen.

A tanulmány további fejezeteiben áttekintjük a Rasch-modell matematikai vonatkozását, illetve a modell egyenletéből levezethető fő tulajdonságait, amelyek egyrészben megkülönböztetik a Rasch modellt a többi IRT modelltől. A tanulmány keretében csak az eredeti, azaz dichotóm adatokra kidolgozott modellel foglalkozunk, nem dichotóm adatokra továbbfejlesztett változataival nem.

A Rasch-modell dichotóm adatokra

Rasch a modell megalkotása során abból indult ki, hogy "a magasabb képességszintű személy nagyobb valószínűség mellett old meg bármely típusú itemet, mint a többi személy és hasonlóan egy item akkor nehezebb, mint a másik, ha bárki nagyobb valószínűséggel oldja meg a másik itemet, mint azt". (*Rasch*, 1960, 117., idézi *Griffin*, 1999)

A modell, mint korábban utaltunk rá, a logisztikus függvényre épít, és a következő matematikai formulát használja az item karakterisztikus görbéjének (4. ábra) meghatározására:

$$p_{ij} = P(x = 1) = \frac{\exp(\theta i - \delta j)}{1 + \exp(\theta i - \delta j)} \qquad (1)$$

ahol x = 1, ha az itemre adott válasz jó és
x = 0, ha rossz,
q_i a személy képességparamétere a vizsgált látens változó képességskáláján,
d_j az itemparaméter (itemnehézség) ugyanazon a skálán.

Az (1) egyenlet a sikeres válaszadás valószínűségét az adott diák képességparaméterének és az item nehézségi indexének függvényében adja meg, pontosabban a kettő különbségének függvényében. Ha a diák képességparamétere azonos az item nehézségi indexével, akkor a helyes válaszadás valószínűsége: 0,5.

Az (1) egyenletet átrendezve:

$$\ln\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \theta i - \delta j \qquad (2)$$

leolvasható, hogy a diák képességparaméterének és az item nehézségének különbsége a helyes és helytelen válaszadás valószínűsége hányadosának (odds) természetes alapú logaritmusa. Ez az oka, hogy a képességszintek és az itemnehézségi paraméterek közös skálájának egysége a logit (log odds unit egy rövidítése).

A továbbiakban áttekintjük a Rasch-modell főbb tulajdonságait, amelyek egyrészt megkülönböztetik az eredeti, dichotóm adatok elemzésére megalkotott Rasch-modellt a többi IRT modelltől, illetve segítenek eredményeink helyes értelmezésében.

A Rasch-modell tulajdonságai

Speciális objektivitás

Az (1) egyenletben bemutatott modell – az IRT modellek közül egyedül – azzal a tulajdonsággal rendelkezik, hogy például két személy összehasonlítása független attól, hogy melyik itemen tesszük azt, illetve két item összehasonlítása független attól, hogy milyen képességszintű személy oldotta meg azokat. Ennek bemutatására a (2) egyenletből indulunk ki és feltételezzük, hogy van két θ_1 és θ_2 képességszintű diákunk, az item, amit megoldanak δ nehézségű. Tegyük fel, hogy p_1 az első személy helyes válaszadásának valószínűsége és p_2 a második személy helyes válaszának valószínűsége.

$$\ln\left(\frac{p_1}{1 - p_1}\right) = \theta_1 - \delta$$

és (3)

$$\ln\left(\frac{p_2}{1 - p_2}\right) = \theta_2 - \delta$$

A két személy helyes és helytelen válaszadása valószínűsége hányadosának (odds) természetes alapú logaritmusának (log odds) különbsége:

$$\ln\left(\frac{p_1}{1-p_1}\right) - \ln\left(\frac{p_2}{1-p_2}\right) = \theta_1 - \delta - (\theta_2 - \delta) = \theta_1 - \theta_2 \qquad (4)$$

A (4) egyenlet alapján a fenti különbség (log odds) független az itemparamétertől és csak a személyek képességparaméterének függvénye. Hasonló átalakítással belátható, hogy az itemek log odds-ának (két itemen adott helyes és helytelen válaszok valószínűségének hányadosának logaritmusa) különbsége pedig a személyek képességparamétereitől független. Mint korábban utaltam rá, ez a féle objektivitás, függetlenség az IRT modellek közül csak a Rasch-modell tulajdonsága.

A képességszintek abszolút helyzetének változása

A (1) egyenlet alapján annak valószínűsége, hogy egy személy egy itemre jó válasz ad függ a személy képességszintje és az item nehézségi szintje közötti különbségtől (θ-δ). A logit skála azonban nem határozza meg a képességszintek és nehézségi indexek abszolút helyét, csak felállítja mind a képességszinteken belül, mind a nehézségi indexeken belül, mind a képességszintek és nehézségi indexek közötti relatív távolságokat. (A helyes válasz valószínűsége attól még nem változik meg, ha a képességszinthez és az itemnehézséghez is hozzáadunk egy konstanst, mivel a kettő különbségének képzésekor az kiesik. *(1. táblázat)*) Ez azt is jelenti, ha például egy skálán van egy 1,2 logitegységes item, egy másik skálán pedig egy 1,5 logitegységes item, a kettőt nem lehet összehasonlítani anélkül, hogy meg ne nézzük, hogyan lettek előállítva a skálák, hova lett a nullpont téve. A két különböző skála egymással történő összehasonlításának problémáját kiküszöbölhetjük, ha a két skálában van valami közös, összekötő elem (diák vagy item), mivel akkor a két mintát közös adatbázisba téve és elemezve összehasonlíthatóvá válnak az eredmények. Egy másik eljárás, hogy mi határozzuk meg bizonyos itemek nehézségi szintjét, lehorgonyozzuk azokat, így a program a többi item nehézségének meghatározásakor azokhoz viszonyít. Ennek az eljárásnak is az a feltétele, hogy legyenek közös itemek.

Azonos diszkriminációs indexek

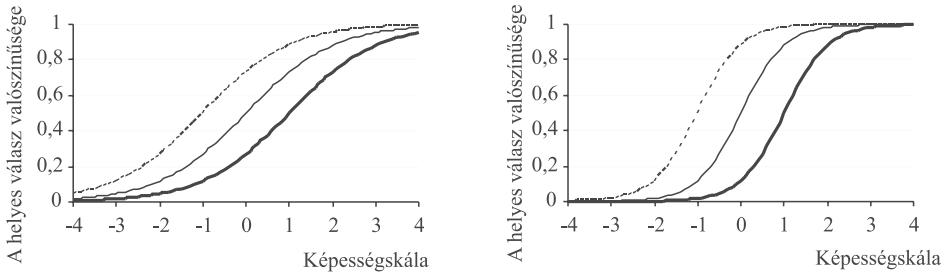
A Rasch modellben egy teszt itemeinek karakterisztikus görbéi elméletileg párhuzamosak, azaz nem metszik egymást és mindegyiknek ugyanaz a meredeksége. *(5. ábra)* A modell ezen tulajdonságát nevezik azonos diszkriminációnak, vagy azonos meredekségnek. Ez alapján a teszt minden egyes iteme diszkrimináló erejének azonosnak kell lenni.

Az abszolút diszkriminációs index változása

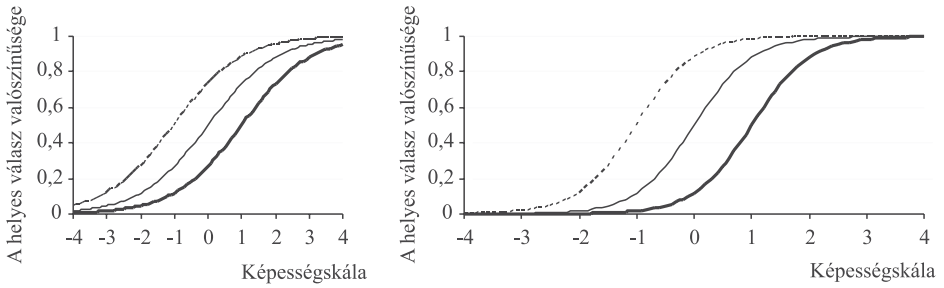
A Rasch-modell ezen tulajdonsága hívja fel a figyelmet arra, hogy nem elegendő egy teszt, illetve a benne lévő itemek modelllilleszkedését (a modelllilleszkedésről részletesebben l. a tanulmány későbbi alfejezetét) megnézni, hanem figyelmet kell fordítani a klaszszikus tesztelméleti reliabilitásmutatóra és az itemek diszkriminációs erejének megfelelő nagyságára. A következőkben egy példán illusztráljuk, hogy elkészíthető olyan teszt, ami tökéletesen illeszkedik a Rasch-modellbe, azaz minden egyes itemének karakterisztikus görbéje párhuzamos, a teszt mégsem jó.

A Rasch-modell a teszt minden itemét azonos meredekségű logisztikus görbével modellálja (képileg), de nem határozza meg a meredekség, azaz a diszkrimináció abszolút nagyságát. Például a *8. ábrá*n két teszt 3–3 itemének karakterisztikus görbéjét modellálljuk. Az első teszt itemeinek karakterisztikus görbéinek meredeksége 1, a második tesztté 2. Ez azt jelenti, hogy a második teszt itemeinek diszkrimináló ereje nagyobb ugyanazon minta esetén. Ha a két teszt itemeit közösen skálázzuk a Rasch modell segítségével, akkor a modell a karakterisztikus görbék meredekségét 1-nek veszi, ezért úgy tűnik, hogy mindkét teszt itemeinek karakterisztikus görbéje párhuzamos lesz (képileg). *(9. ábra)* Azonban a jobban diszkrimináló itemek jobban diszkriminálják, széthúzzák a diákok ké-

pességparaméter-értékeit a képességskálán. Elegendő csak az abszcissza tengely osztáspontjainak változására nézni.



8. ábra. Különböző diszkriminációs erővel bíró itemek karakterisztkus görbéi (Wu, 2006a alapján)



9. ábra. A 8. ábra itemeinek görbéi Rasch skálázás után (Wu, 2006a alapján)

A Rasch-modellhez való illeszkedéshez visszatérve, mindkét teszt itemei ugyanolyan jól illeszkednek a modellhez, de ha a két tesztet összerakjuk egy tesztté, már lesznek a modellbe nem illeszkedő itemek is. Sarkítva: lehet olyan tesztet készíteni, amelyik csak olyan itemekből áll, amelyek karakterisztikus görbéi nagyon laposak. Ez azt jelenti, hogy minden egyes itemet a diákok találgatással oldanak meg, azaz az itemek nem képesek a különböző képességű diákok megkülönböztetésére, ugyanúgy teljesítenek a teszten az alacsonyabb és a jó képességű diákok is. Ennek ellenére a teszt jól illeszkedik a modellbe, mivel a görbék egymással párhuzamosak, azonos diszkrimináló erejűek. Ez az oka, hogy nem elegendő csak az illeszkedést megnézni, fontos a reliabilitásmutató is, aminek alacsony értéke az előbbi példa esetében rávilágítana arra, hogy nem jó a tesztünk. Egy rossz tesztet nem tesz jóvá a Rasch-modell, csak az eredmények elemzéséhez más eszközöket is kínál.

A logitegységnek nincs abszolút hossza

A Rasch-modell előbbi tulajdonsága során rámutattunk arra, hogy a logitegységnek nincs abszolút hossza. Ez azt jelenti, hogy tesztfüggő, milyen távol van egymástól két ember képességparamétere a képességskálán. Egy magasabb diszkrimináló erővel rendelkező teszt jobban széthúzza, jobban diszkriminálja a személyeket, mint egy, az adott mintát kevésbé diszkrimináló feladatlap, még akkor is, ha az esetleg jobban illeszkedik a Rasch-modellhez (l. előbbi szélsőséges példát). A valóságban egy teszt itemei sohasem rendelkeznek azonos diszkrimináló erővel (általában a feleletalkotó kérdések például jobban diszkriminálnak, mint a feleletválasztó kérdések).

A Rasch-modell fő alkalmazási területe

Abban az esetben, ha egy minta minden egyes tagja ugyanazt a tesztet oldotta meg, az eredmények pontos elemzéséhez nincs szükség Rasch-modellre, a teljesítmények nyers-

pont-értékei elegendőek a megfelelő statisztikai számolások elvégzéséhez és értelmezéséhez. Ha kiszámolnánk ebben az esetben a nyerspont-értékek és a képességszintek közötti korrelációt, 1-hez közeli korrelációs együtthatót kapnánk.

Ha viszont a vizsgálatot úgy építjük fel, hogy különböző diákok, különböző, de horgony itemeket tartalmazó teszteket oldanak meg, aminek következtében a vizsgálatban szereplő nem minden itemet old meg minden diák, az eredmények elemzéséhez és az egyes, különböző tesztet megírt diákcsoportok közötti összehasonlító vizsgálatok elvégzéséhez már Rasch-modellre van szükség. Ebben az esetben, ha a megoldott itemek halmazától független meghatározást szeretnénk, a diákok képességszintjének meghatározásához nem elegendőek a nyerspont-értékek.

Az adatok illeszkedése a Rasch-modellhez (model fit)

A Rasch-modell fent felsorolt tulajdonságai abban az esetben érvényesek, ha az empirikus vizsgálat mérőeszköze illeszkedik a Rasch-modellbe. A Rasch-modell szempontjából annál jobb a mérőeszköz: minél diszkriminálóbb itemeket tartalmaz; az itemek diszkrimináló ereje közel azonos; viszont nehézségi indexük eltérő, hogy az itemek nehézségi skálája lefedje a diákok képességeloszlását; az empirikus item karakterisztikus görbék közel vannak az elméleti görbékhez.

A Rasch-modell a helyes válasz valószínűséget az item nehézsége (δ) és a személy képességparamétere (θ) alapján határozza meg - l. (1) egyenlet. Ebből adódóan, ha egy itemen a helyes válasz valószínűségét más is befolyásolja, mint az item nehézsége és a személy képességparamétere, akkor sérül a Rasch-modell alkalmazhatósága. Néhány tényező, ami rontja az itemek modellilleszkedését:

– Találgatás – Főleg magas nehézségi indexű, azaz nagyon nehéz feleletválasztós itemek esetén fordul elő. Általában a feleletalkotó kérdések diszkrimináló ereje jobb, mint a feleletválasztós kérdéseké. (*Wu*, 2006b)

– Itemfüggőség – Ha egy item helyes megoldásához egy másik itemen adott választ kell felhasználni (erős függőség), vagy a kontextus összeköti az itemeket (könnyű függőség) (*Wu*, 2006a).

– Különböző itemműködés – DIF (Differential Item Functioning). A minta különböző kohortjai máshogy válaszolnak a kérdésre, például a fiúk általában jobban válaszolnak a focival kapcsolatos kérdésekre, mint a lányok. (Erre példát lásd *Molnár,* 2006 tanulmányában.)

– Többdimenzionalitás - Ha egy item mást, más látens képességet mér, mint a többi item. (Például, ha egy matematika item mind a fogalmi értést, mind a számolási képességet méri. Ez a két látens változó pedig személyenként változhat, valaki az egyikben jobb, valaki a másikban.)

Azt, hogy alkalmazhatjuk-e az adataink elemzésére a Rasch modellt, az illeszkedésvizsgálat (fit statistics) mutatja meg. A kutatók számos matematikai modellt dolgoztak ki (*Write és Masters,* 1982) a modellilleszkedés tekintetében, amelyek a fent említett – a Rasch-modell alkalmazhatóságát befolyásoló – tényezők közül minden esetben csak egyet vizsgálnak (pl.: a diszkriminációs indexek egyezése vagy az érvényesül-e az egydimenzionalitás), csak egy feltétel teljesülését ellenőrzik. Ezt azért lényeges megemlíteni, mert általában már egy illeszkedésvizsgálat után következtetéseket vonunk le, holott lehet, hogy ha egy másik modellt használtunk volna, aszerint nem ugyanazt az eredményt kaptuk volna. Az illeszkedésvizsgálatok között e tanulmány keretében a maradék alapú (residual based) illeszkedésvizsgálat főbb tulajdonságait mutatjuk be, mivel többek között ezt használja az általunk használt ConQuest (*Wu, Adams* és *Wilson*, 1998) szoftver. Ezen túl a Quest (*Adams és Khoo*, 1996), a Winsteps (*Linacre és Write*, 2000) és RUMM (2001) is (idézi *Wu*, 2006b)

A residuális alapú illeszkedésvizsgálat (residual based fit statistics)

Az illeszkedés nagyságát a programok két lépcsőben számolják ki. Első lépésként meghatározzák a személy képességparaméterét és az item nehézségi mutatóját, majd azokból kiszámolják mind a személy, mind az item illeszkedését. (Előbbi jelentését l. *Molnár,* 2005) A program a számoláshoz első lépésben képez egy mátrixot, ami minden egyes diák minden vizsgált itemen elért eredményét (később x_{ij}) (0 vagy 1) tartalmazza. Ebből generál egy olyan mátrixot, ahol az egyes helyeken már a helyes válasz elvárt valószínűsége [később $E(x_{ij})$ – Rasch Expected Response Probabilities (*Bond és Fox*, 2001)] áll. A két mátrixot egymásból kivonva ($y_{ij} = x_{ij} - E(x_{ij})$) megkapjuk a harmadik mátrix elemeit (response residual). Ezek után minden egyes elemet sztenderdizálni kell. A sztenderdizált modell egyenlete:

$$z_{ij} = \frac{x_{ij} - E(x_{ij})}{\sqrt{Var(x_{ij})}} \qquad (5)$$

ahol x_{ij}: i személy j itemen megfigyelt eredménye,
$E(x_{ij})$: i személy j itemen történő helyes válaszának valószínűsége. (Wu, 2006b)

A dichotóm Rasch-modell esetén a $E(x_{ij}) = p_{ij}$ és $Var(x_{ij}) = p_{ij}(1 - p_{ij})$, azaz $Var(x_{ij}) = I_{ij}$ *(1. táblázat).* Az ezen a módon kiszámolt maradékok képezik a modell illeszkedésvizsgálatának alapját. A programok (pl.: Quest) az egyes itemek modellilleszkedését általában grafikusan jelenítik meg. (erre példát ld. *Molnár,* 2005, *Molnár és Józsa*, 2006)

A j-edik item illeszkedési indexének (fit index) meghatározásához a program először négyzetre emeli, majd nullától i-ig összeadja a négyzetre emelt maradékokat (z_{ij}), míg az i-edik személy illeszkedési indexének meghatározásához a maradékok (z_{ij}) négyzetre emelése után azokat nullától j-ig adja össze. (*Write és Masters*, 1982) Az item illeszkedésre Write és Masters (1982) két statisztikai módszert javasolt: egy súlyozott (más néven infit) [unweigted mean-square (MNSQ)] és egy súlyozatlan (más néven outfit) (weigted MNSQ) értéket.

Az outfit a négyzetre emelt sztenderdizált maradékok hagyományos összeadásán alapul:

$$\text{SúlyozatlanMNSQ} = \frac{\sum_i z_{ij}^2}{n} = \frac{1}{n}\sum_i \frac{(x_{ij} - E(x_{ij}))^2}{Var(x_{ij})} \qquad (6)$$

ahol n: a válaszadók száma.
Az outfittel szemben, ahol minden egyes súlya azonos: 1, az infit a következőképpen definiált (Wu, 2006b):

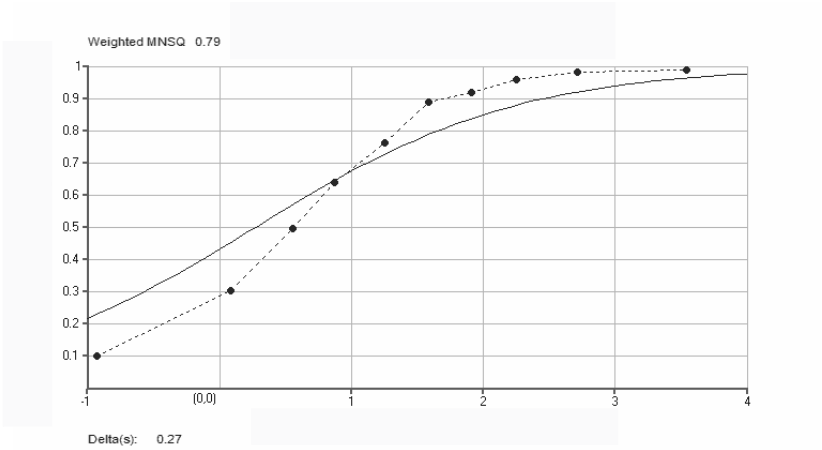
$$\text{SúlyozottMNSQ} = \frac{\sum_i z_{ij}^2 Var(x_{ij})}{\sum_i Var(x_{ij})} = \frac{\sum_i (x_{ij} - E(x_{ij}))^2}{\sum_i Var(x_{ij})} \qquad (7)$$

Ebben az esetben minden egyes z_{ij}^2 súlyozott $Var(x_{ij})$-vel és a nevezőt a súlyok összege adja. Ha elvégeznénk az összeadásokat, belátható, hogy mind az infit, mind az outfit érték 1-hez tart, azaz, ha az adatok illeszkednek a modellhez, akkor az MNSQ (infit és outfit esetén is) értéke 1. Ebben az esetben az itemek diszkrimináló ereje közel azonos, azaz teljesül a Rasch modell szempontjából jó mérőeszköz ismérveinek 2. pontja (l. fent). Ha ez nem teljesül és az MNSQ értéke távol esik 1-től, akkor az adott item nem illeszkedik a többi item által alkotott modellbe. Ha az item nehézségi indexe és a személy képességparamétere közel van egymáshoz, akkor a $Var(x_{ij})$ értéke relatív magas, azaz több információval szolgál, mint azok az itemek, amelyek nehézségi indexe jóval alacsonyabb, vagy jóval magasabb, mint a személy képességparamétere illetve azok a személyek, akik képességparamétere távol van az adott item nehézségi indexétől. Ezekben az esetekben kisebb súllyal számol a modell. Vajon 1-től milyen mértékű eltérést fogadunk még el, mikor mond-

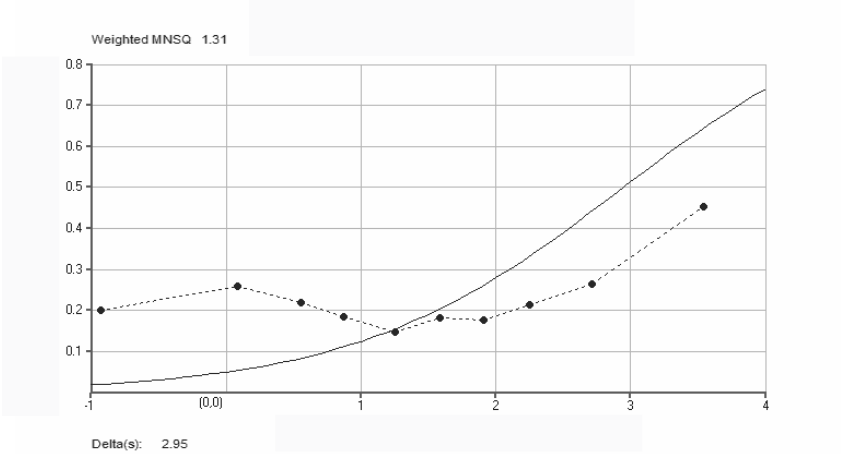
hatjuk még, hogy az adatok illeszkednek a modellhez, és mikortól beszélünk alul-, vagy túlilleszkedésről?

Első lépésként nézzük meg, hogy mit jellemez az MNSQ értéke, mi az a tulajdonság, aminek teljesülését ellenőrzi? A (6) egyenletben definiált statisztikai módszer azt ellenőrzi, hogy az egyes itemek karakterisztikus görbéinek meredeksége azonos-e – a Rasch-modell abból a feltételezésből indul ki, hogy az itemek karakterisztikus görbéi párhuzamosak. Ha az empirikus adatokból felépített item karakterisztikus görbe meredekebb, mint az elméleti görbe, akkor az MNSQ értéke kisebb, mint 1, de ha a tapasztalt item karakterisztikus görbe laposabb, mint az elméleti görbe, akkor az MNSQ értéke nagyobb, mint 1. A *10*. és *11*. *ábra* mindkét esetre bemutat egy példát.

Az MNSQ értéke nem arról ad információt, hogy az empirikus görbe egyes pontjai milyen távol vannak az elvárt görbétől, hanem, hogy az empirikus item karakterisztikus görbe átlagos meredeksége mennyire közelíti az elvárt görbe átlagos meredekségét, azaz mennyire azonos a diszkrimináló erejük. Az empirikus görbe pontjainak elvárt görbétől való távolságát sokkal inkább a klasszikus tesztelméletből is ismert reliabilitás és diszkriminációs index jellemzi. Egy jól mérő, jól viselkedő item (magas reliabilitás és diszkriminációs erő) esetében, ha az MNSQ értéke 1, akkor az elméleti és az empirikus görbe egymásra fekszik.



10. ábra. Példa egy, az elvártnál meredekebb karakterisztikus görbére (MNSQ=0,79)



11. ábra. Példa egy, az elvártnál laposabb karakterisztikus görbére (MNSQ=1,31)

Az MNSQ fenti értelmezése alapján belátható, hogy ez a típusú illeszkedésvizsgálat nem arról ad információt, hogy általánosságban jó-e, vagy rossz-e az item, hanem arról, hogy mennyire illeszkedik a többi item közé. Ebből adódóan, mint a Rasch-modell főbb tulajdonságai között is jeleztük már, nincs előre meghatározott abszolút jó diszkriminációs index. A fenti kérdésre válaszolva – az MNSQ értékében 1-től milyen mértékű eltérés fogadható el –, azok az itemek, amelyek MNSQ értékei kevesebb mint 1 + 2 (sztenderd hiba) távolságban vannak. (*Perline, Write és Wainer,* 1979) Az (5) egyenletből levezethető, hogy a sztenderd hiba = $\sqrt{(2l)/(n-1)(l-1)}$, ahol n a minta elemszámát, l az itemek számát jelenti, ami $\sqrt{2/n}$-hez tart. Azaz, ha alacsony a minta elemszáma, akkor az itemek MNSQ értékei jobban ingadoznak 1 körül, mint ha magas a minta elemszáma. Például egy 200 fős minta esetén az elfogadható MNSQ értékek 0,8 és 1,2 között ingadoznak, egy 2000 fős minta esetén pedig 0,94 és 1,06 között. Ha az MNSQ értékek elfogadható értékeinek sávja függ a minta elemszámától, akkor nem tudunk egy előre meghatározott elfogadási sávot mondani, hanem minden egyes esetben külön mérlegelni kell. (*Wu,* 2006b)

Az MNSQ értékek egy további transzformációját is javasolta Write és Masters, ahol a transzformáció egyenlete figyelembe veszi az MNSQ értékek átlagát és szórását. Az így nyert értékeket nevezzük t értékeknek, amelyek már közel normál eloszlásúak (átlag = 0, szórás = 1). Ebből adódóan a t értékek elfogadható intervalluma 95 százalékos konfidencia szinten: (-1,96; 1,96). Az MNSQ értékek ezen transzformációja látszólag megoldotta a mintafüggőség problémáját, azonban a valós életben nem létezik olyan item, ami tökéletesen illeszkedne a modellhez. Ebből adódóan, a magas minta elemszám a kicsi eltéréseket is felnagyítja, például egy 300 fős mintán illeszkedőnek tűnő itemek egy 15000 fős mintán már nem illeszkednek a modellhez. Ez azt a dilemmát okozza, hogy az MNSQ értékek alapján annál jobban illeszkednek az itemek, minél nagyobb a minta elemszáma, viszont a t értékeknél az alacsonyabb minta elemszám esetén figyelhetünk meg jobb illeszkedést.

A problémát feloldani úgy lehet, hogy egyedül az illeszkedésvizsgálatok eredménye alapján nem törlünk ki itemet, hanem, az illeszkedésvizsgálatok eredményét, mint diagnózist értelmezzük, ami rávilágít az esetleges problémás itemekre. Ezeket több oldalról meg kell vizsgálni és utána meghozni a döntést, hogy kihagyjuk-e a későbbi felmérésekből vagy nem. *(1)*

Jegyzet

(1) A tanulmány a T 046659PSP OTKA kutatási program, az Oktatáselméleti Kutatócsoport és az SZTE MTA Képességkutató Csoport keretében készült. A tanulmány írása idején a szerző Bolyai János Kutatási Ösztöndíjban részesült.

Irodalom

Adams, R. J. – Khoo, S. (1996):*Quest: The interactive test analysis system*. ACER, Camberwell.

Adema, J.J. – Gademann, A.J.R.M. (1992): Computerized Test Construction. In Wilson, M. (szerk.): *Objective Measurement. Theory into practice*. Ablex Publishing Corporation, Norwood, New Jersey. 261–273.

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

Griffin, P. (1999): *Item Response Modelling: An introduction to the Rasch Model*. Assessment Research Centre Faculty of Education, The University of Melbourne.

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

Linacre, J. M. – Write, B. D. (2000): WINSTEPS: *A Rasch computer program*. MESA Press, Chicago.

Molnár Gyöngyvér (2003): Az ismeretek alkalmazásának vizsgálata modern tesztelméleti eszközökkel. *Magyar Pedagógia*, 4. 423–446.

Molnár Gyöngyvér (2004): Hátrányos helyzetű diákok problémamegoldó gondolkodásának fejlettsége. *Magyar Pedagógia*, 3. 319–338.

Molnár Gyöngyvér (2005): Az objektív mérés megvalósításának lehetősége: a Rasch-modell. *Iskolakultúra*, 3. 71–80.

Molnár Gyöngyvér (2006): 2–11. évfolyamos diákok olvasási képességnek fejlettsége (elemzések a Rasch modell alkalmazásával). Kézirat.

Molnár Gyöngyvér – Józsa Krisztián (2006): Az olvasási képesség értékelésének tesztelméleti megközelítései. In: Józsa Krisztián (szerk.): *Az*

olvasási képesség fejlődése és fejlesztése. Dinasztia Tankönyvkiadó, Budapest. (megjelenés alatt)
Perline, R. – Wright, B. D. – Wainer, H. (1979): The Rasch Model as Additive Conjoint Measurement. *Applied Psychological Measurement*, 2. 237–255.
Rasch, G. (1960): *Probabilistic models for some intelligence and attainment tests.* Danish Institute for Educational Research, Copenhagen.
RUMM Laboratory (2001): *Rasch Unified Measurement Models*. Perth.
Write, B. D. – Masters, G. N. (1982): *Rating Scale Analysis.* MESA press, Chicago.
Write, B. D. – Stone, M. H. (1979): *Best Test Design*. MESA press, Chicago.
Wu, M. (2006a): *PISA Training Workshop: Application of Item Response Theory (IRT) to PISA (ConQuest).* Hong Kong PISA Centre, Hong Kong.
Wu, M. (2006b): *How Well Do the Data Fit the Model?* Kézirat.
Wu, M. – Adams, R. J. – Wilson, M. R. (1998): *ACER ConQuest. Generalised Item Response Modelling Software*. ACER Press, Australia.