SZABÓ GÁBOR

# Ponthatár vagy képességszint?

## Az IRT-alapú teljesítményszint-minimumok megállapításának gyakorlati kérdései a nyelvtudásmérésben

### 1. Bevezetés

Bármely pszichológiai típusú mérésről is legyen szó, általában kulcskérdésnek számít a minimálisan elfogadható teljesítmény szintjének meghatározása, vagy közkeletűbb kifejezéssel élve a „ponthatárok" megállapítása. Ez természetesen nem meglepő, hiszen gyakorta nagyon komoly, a vizsgázó számára akár hosszú távú következményekkel is járó döntéseknek szolgáltatnak alapot a ponthatárok, illetve a megfelelési minimumok. Az effajta döntések esetében igen fontos, hogy a ponthatárok valós különbségeket tükrözzenek, s hogy kellő megalapozottsággal bírjanak.

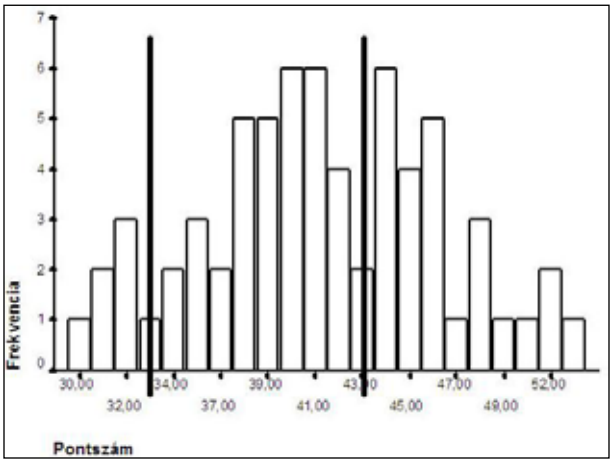
Mindez igaz a nyelvtudás mérésének az esetében is, és a téma különösen releváns napjainkban, amikor a vonatkozó kormányrendelet 2005. évi módosítását követően valamennyi Magyarországon akkreditált nyelvvizsga-rendszernek 2006 szeptemberéig az Európa Tanács által kidolgozott Közös Európai Referenciakeret (KER 2002) szintjeihez kellett igazítania vizsgaszintjeit. Ennek kapcsán azt is demonstrálni kellett, hogy a megállapított teljesítményminimumok híven tükrözik a KER elvárásait.

Ebben az írásban arra teszünk kísérletet, hogy a hagyományos teljesítményszint minimum-meghatározási eljárásaival szemben egy alternatív, több szempontból hatékonyabb módszert írjunk le. Először rövid áttekintést adunk a tradicionális megközelítés adta lehetőségekről, illetve ezek korlátairól, majd bemutatjuk a modern, probabilisztikus tesztelmélet adta lehetőségeket. Ezután kitérünk az alternatív eljárások gyakorlati előnyeire és hátrányaira, végül megkíséreljük levonni mindezekből a következtetéseket.

### 2. A teljesítményminimumok hagyományos megállapítása

A hagyományos eljárások vizsgálata kapcsán először is célszerű különbséget tennünk a norma-orientált és kritérium-orientált tesztelés között. Előbbi esetében a vizsgázók teljesítménye kizárólag a többi vizsgázó teljesítményének függvényében kerül értékelésre, míg utóbbi kapcsán a vizsgázók eredményeit külső, előre definiált elvárásokhoz viszonyítva vizsgálhatjuk meg (McNamara 2000: 62–64).

A norma-orientált mérések esetében a megfelelési minimumokat gyakorta a ponteloszlás valamilyen grafikus megjelenítésének segítségével határozzák meg. Erre ad példát az 1. ábra.




***1. ábra.*** **Ponteloszlási oszlopdiagram lehetséges ponthatárokkal (normaorientált)**

Az ábrán egy lehetséges ponteloszlás látható, oszlopdiagram formájában megjelenítve. Amint az világosan látszik, több helyen eloszlási „völgyeket" találunk, s gyakran pontosan ezek jelentik a lehetséges ponthatárok meghatározásához a vizuális segítséget. Az ábrán két lehetséges ponthatárt két függőleges vonal jelez. Természetesen a ponthatárok meghatározása számos tényező függvényében történhet, s az sem előírás, hogy a ponthatároknak eloszlási völgyben kell lenniük. De akár a vizsgázók egyszerű rangsorba állításával, akár a fenti eljárással határozzák meg a megfelelési minimumot, az végső soron csakis a vizsgázók egymáshoz viszonyított teljesítményére épül (Alderson, Clapham és Wall 1995: 156–157).

Ehhez képest a kritérium-orientált teljesítmény-minimumok megállapítása más eljáráson alapul, melynek kapcsán különféle standardizálási eljárások valamelyike, esetleg ezek közül több is alkalmazásra kerül. A standardizálás tesztközpontú, illetve vizsgázóközpontú módszerekkel történhet, annak függvényében, hogy a viszonyítási alap az itemek[1] nehézsége vagy a vizsgázók képessége-e (Manual 2009: 60). Gyakori módja a standardizálásnak, hogy az adott tesztet szakértők megoldják, és általában az Angoff-eljárás valamely változatát alkalmazva (Angoff 1971; Livingstone és Zieky 1982; Hambleton és Plake 1995) eldöntik, mely itemeket milyen valószínűséggel válaszol meg helyesen egy olyan vizsgázó, akinek teljesítménye éppen megüti a megfelelési mércét. A szakértői vélemények összegzése alapján aztán megállapítható, milyen pontszámot ér el a fentebb már említett vizsgázó, s ez lesz egyben a kritériumorientált ponthatár is.

Az esetek többségében azonban valójában a kritérium-orientált tesztek esetében sem teljesen független a ponthatárok megállapítása a vizsgázók tényleges teljesítményétől. Ennek egyik oka abban keresendő, hogy egy még oly szakszerűen alkalmazott

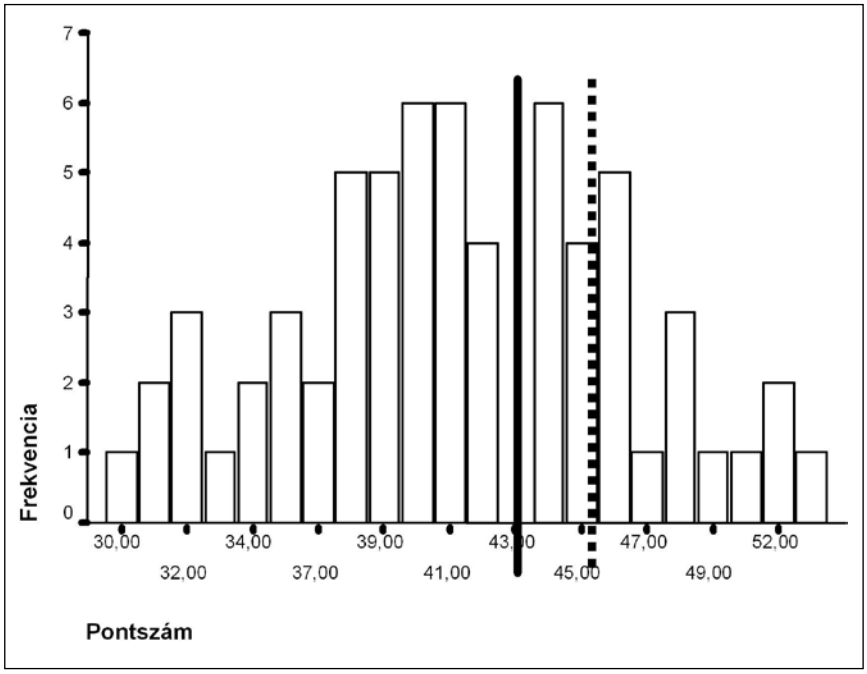
[1] Az item (a magyar szakterminológiában használt alternatív kifejezések: ütem, tétel, feladattétel, elem): egy teszt legkisebb olyan alkotóeleme, amelyre a vizsgázó specifikus módon válaszol, és erre pontot vagy pontokat kap; például egy feleletválasztós kérdés a válaszopciókkal együtt, egy kitöltendő rubrika egy táblázatban stb.

Angoff-eljárás esetében is előfordul, hogy a szekértők különböző következtetésekre jutnak, illetve az átlagolt szakértői vélekedés néhány pontos különbséghez vezethet. Márpedig akár egyetlen pont is egy adott vizsgázó esetében döntő fontosságú lehet.

A másik ok abban rejlik, hogy számos vizsga – különösen nyelvvizsga – esetében a megfelelési minimum már a vizsgafeladatok megalkotása előtt eldöntött és ismert tény. Természetesen a feladatírók igyekeznek ennek függvényében tevékenykedni és azonos nehézségi szintű feladatokat alkotni a vizsga különböző verzióihoz, ám 100%-ig azonos tesztváltozatokat igen nehéz előállítani, s azt, hogy a ponthatár tekintetében egyetlen pont különbség se legyen, gyakorlatilag nem lehet garantálni.

Mindebből következik, hogy a kritérium-orientált vizsgák esetében is szóhoz jutnak a ponteloszlási adatok, gyakorta oly módon, ahogyan az a 2. ábrán látható.

Az ábrán ismét egy lehetséges ponteloszlást látunk. A pontozott vonal jelöli azt a pontszámot, amely a kritérium-orientált érték az előzetes szakértői vélemények, illetve a teszt korábbi verziói alapján. Az is jól látszik azonban, hogy a tényleges teljesítmények között nagyobb, valósabb különbség tehető, ha az eredetileg megállapított 45 pont helyett az ettől csak kevéssé eltérő 43 pontnál húzzuk meg a ponthatárt; ezt az ábrán folyamatos vonal jelöli. Számos vizsgarendszer a fentebb leírt bizonytalansági tényezők miatt alkalmazza is ezt az utólagos „finomhangolást”, amely – és ezt érdemes kiemelni – nem csökkenti, hanem éppen növeli a kritérium-orientált eljárás szakszerűségét, hiszen kompenzálja a szakértők, illetve a tesztverziók különbözősége miatt megjelenő pontatlanságot.




***2. ábra.*** **Ponteloszlási oszlopdiagram lehetséges ponthatárokkal (kritériumorientált)**

## 3. A modern tesztelmélet adta lehetőségek

A modern tesztelmélet (Item Response Theory – IRT) alkalmazása a nyelvtudásmérés területén még viszonylag rövid múltra tekinthet vissza. Ám ez alatt az idő alatt is világossá vált, milyen sokrétűen alkalmazható a klasszikus tesztelméleti megközelítés kiegészítéseként.

A teljesítmény-minimumok meghatározása kapcsán a probabilisztikus tesztelmélet két alapvető előnyt kínál. Egyrészt az egyes vizsgázókat illetően a pontszámok helyett képességindexek állapíthatók meg, s a minimum-követelményeket is a képességindexek formájában lehet megadni. Másrészt, ezzel összefüggésben, a teljesítmény-minimumok objektívebb formában jeleníthetők meg, s így az összehasonlítás tágabb keretek között valósulhat meg.

Az alábbiakban vizsgáljuk meg e két előnyös tényezőt közelebbről. A modern tesztelméleti megközelítés az itemek és a vizsgázók közös nehézségi-képességi kontinuumon történő elhelyezésére épül. Jelen vizsgálódásunk szempontjából ennek azért van jelentősége, mert ily módon a vizsgázók teljesítményét sem egyszerűen a helyes válaszok számának összege, a pontszám adja (mely a nyerspontok esetleges konverziója után sem ad ennél több információt), hanem egy viszonylag bonyolult, valószínűségszámítási matematikai modell alapján megállapított képességindex (Bachman 2004: 142).

A modell részletes leírásától terjedelmi okokból itt el kell tekintenünk, ám a szakirodalomban számos forrás részletesen leírja (például Hambleton és Swaminathan 1985; Hambleton, Swaminathan és Rogers 1991; Hulin, Drasgow és Parsons 1983). Ezen a helyen azt szükséges megállapítani, hogy a modell abból indul ki: egy adott nehézségi szintű itemet egy azzal megegyező képességszintű vizsgázó 50% valószínűséggel old meg helyesen, illetve helytelenül. Magasabb képességszint esetén a helyes, alacsonyabb képességszint esetén a helytelen válasz valószínűsége nő, s a valószínűség a nehézségi index és a képességindex közötti különbség függvénye. Ily módon egy vizsgázónak a teszt valamennyi itemjére adott válaszai alapján megállapított képességindexe nem csupán a helyes válaszokra, illetve azok számára épül, de az itemek nehézségi szintjét is tekintetbe veszi. Ennek pedig abban áll a jelentősége, hogy míg egy tesztpontszám nagyon különböző válaszmintákat is takarhat, s nem ad információt arról, *melyek* voltak a vizsgázó által helyesen megválaszolt itemek, a képességindex alapján a vizsgázó tényleges teljesítményéről árnyaltabb, realisztikusabb képet kapunk. A képességindex értékét az ún. *logit* pontszámmal szokás megadni (McNamara 1996: 164–165).

A nyers vagy akár konvertált teszt-pontszámokhoz képest a logit értékek további előnye, hogy míg a teszt-pontszámok ordinális skálán helyezkednek el, addig a logit pontok intervallum skálát alkotnak. Ennek oka megint csak az, hogy nem tudható, két azonos teszt-pontszám azonos tudásszintet takar-e. Ami azt illeti, az esetek többségében inkább azt kellene feltételeznünk, hogy nem azonosat, hiszen a teszt egyes itemjei általában szándékoltan különböző nehézségűek, ezáltal az elért pontszám jelentése attól függ, mely itemeket válaszolta meg helyesen a vizsgázó. Vagyis nem tudható, hogy a 20 és 21 pontos teljesítmény között ugyanakkora-e a különbség, mint a 21 és 22 pontos között.

Ehhez képest a logit pontok, mivel nem a nyerspontokra épülnek és mert megállapításukkor az itemek nehézsége is szerepet játszik, a vizsgázók egymáshoz viszonyí-

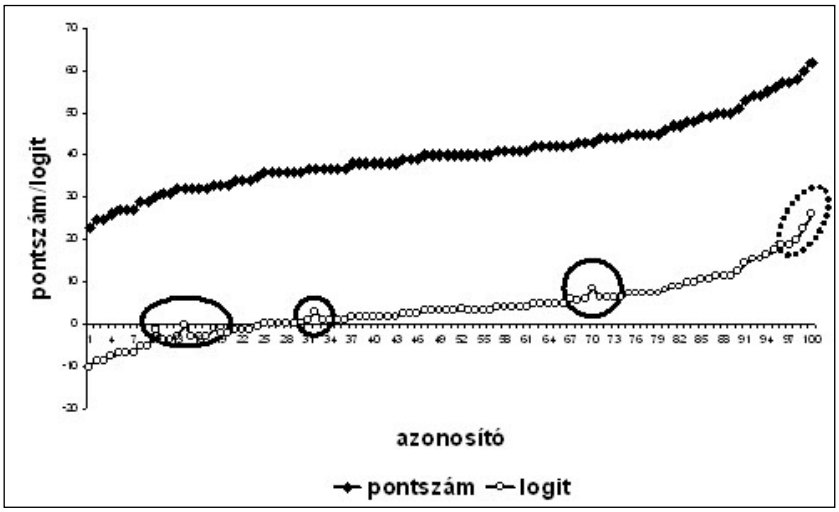
tott teljesítményéről is hívebb képet adnak, s a vizsgázók közti különbségek nagysága is jobban megragadható, értelmezhető.

A gyakorlatban ez azt jelenti, hogy két azonos pontszámmal rendelkező vizsgázó képességindexe eltérést mutathat, amennyiben más nehézségi fokú itemekből válaszoltak meg helyesen ugyanannyit.

A fentieket jól érzékelteti a 3. ábrán szereplő két eloszlási grafikon. Az ábrán valós teszteredmények láthatóak. A Pécsi Tudományegyetem angol szakos képzésében a nyelvi alapvizsga nyelvi-nyelvhasználati komponense 2005-ös verziójának eredményei alapján az elért nyerspontok, illetve az ezeknek megfeleltethető képességindexek két görbét rajzolnak ki. A vízszintes tengelyen a vizsgázók szerepelnek az elért nyerspont sorrendjében, míg a függőleges tengelyen a pontszám, illetve a logit érték szerepel. Mint látható, a pontszámok és a logit pontok eloszlása nagyjából hasonló képet mutat a tendenciákat tekintve. A két érték közötti számszerű különbség nem lényeges, a logit pontszám kezdőértéke tetszés szerint szabályozható. Ami figyelmet érdemel az általános hasonlóságon túl, az a körökkel jelzett öt képességindex-érték. Amint az megfigyelhető, a kiemelt logit értékekhez tartozó nyerspontszámok nem különböznek a velük „szomszédos" pontszámoktól, ám a logit pontokban különbség mutatkozik, méghozzá helyenként akkora, ami „lefordítva" több nyerspontnyi differenciát jelent.

Mindez jól érzékelteti, hogy a nyerspontok egyezése nem jelenti feltételül a képességindexek egyezését is. Ugyancsak figyelemre méltó az ábra jobb oldalán a pontozott ellipszissel jelölt három logit érték. Ha ezeket összehasonlítjuk a hozzájuk kapcsolódó nyerspontszámokkal, azt láthatjuk, hogy a logit pontok által kirajzolt vonal meredekebben emelkedik, mint a nyerspontok által maghatározott vonal. Más szóval, a képességindex tekintetében nagyobb különbségek állapíthatóak meg a három jelzett vizsgázó között, mint amit a nyerspontok alapján detektálhatunk.

A fentiek alapján világosan látszik tehát, hogy pontszámok helyett képességindex-értékeket használva pontosabb képet alkothatunk a vizsgázók teljesítményéről. Ebből




***3. ábra.*** **Pontszám és képességindex-eloszlási grafikon**

az is következik, hogy a teljesítmény-minimumokat is meg lehet adni logit pontok formájában, ami azt jelentheti, hogy a ponthatároknál hatékonyabb szűrőt állíthatunk fel. A hatékonyság azért lesz jobb, mert maga a mérés válik pontosabbá. Ha a 3. ábrán körrel jelzett logit értékekhez tartozó vizsgázók nyerspontszáma éppen a ponthatár alá esne (s ez különösen kritériumorientált tesztek esetében lényeges), ők méltatlanul hullanának ki a szűrőn, hiszen a logit pontszám alapján látható, hogy tényleges képességszintjük magasabb, mint a velük azonos nyerspontszámot elért vizsgázóké. Ha viszont a minimumszint logit pontértékben kerül meghatározásra, nem éri méltánytalanság a fent említett vizsgázókat. Ráadásul a minimumszint ilyetén definiálása más előnyökkel is járhat.

A nyelvvizsgarendszerek többsége előre közzéteszi a leendő tesztek „ponthatárait". Ezzel tulajdonképpen azt vállalja, hogy a különböző tesztverziók pontosan egyforma nehézségi szintűek lesznek, és pontosan azonos pontszám jelzi ugyanazt a nyelvtudás-szintet. Amint erről korábban már szóltunk, ez gyakorlatilag nem kivitelezhető, s így az alapvetően kritérium-orientált vizsgák norma-orientált módszereket kénytelenek alkalmazni. Ám mindez kivédhető, ha a megfelelési minimumot nem pontszámban vagy ebből származó százalékos értékben definiáljuk, hanem képességindex-értékben. Ebben az esetben a megfeleléshez szükséges pontszám ugyan eltérő lehet a különböző tesztverziókban, ám a tényleges képességindex-határ ugyanaz marad.

Minderre viszont csak abban az esetben van lehetőség, ha a különböző tesztverziók valamilyen technikával össze vannak kapcsolva, vagyis ún. horgony itemeket tartalmaznak. Ennek előnyeiről és hátrányairól a későbbiekben még lesz szó.

A fentieket figyelembe véve vizsgáljuk most meg, milyen pozitív és negatív hozadéka lehet a képességindex-értékek gyakorlati alkalmazásának.

## 4. Praktikus előnyök és hátrányok

A fentiekben láthattuk, hogy a modern tesztelméleten alapuló képesség logit értékek használata több szempontból hasznos lehet. Ám a pozitívumok mellett potenciális nehézségekről is szót kell ejteni. Mielőtt azonban ezeket sorra vennénk, tekintsük át még egyszer azokat az egyértelműen előnyös aspektusokat, amelyek az IRT alapú megközelítést jellemzik.

Először is, mint azt fentebb már részletesen megvizsgáltuk, a logit értékek alkalmazásával pontosabb és igazságosabb lehet a teljesítmény-minimumok meghatározása. Pontosabb, mert a vizsgázók közötti tényleges különbségeket megmutató intervallum skála adatok használatára nyílik mód, és igazságosabb, mert a ponthatárokhoz képest a valós képességszintet jobban tükröző megfelelési minimumokat lehet meghatározni.

Ez utóbbi szempont napjainkban különös aktualitást is kap azáltal, hogy a nyelvtudásmérés európai, sőt Európán kívüli színterein is egyre nagyobb hangsúllyal jelenik meg a különféle vizsgarendszereknek a Közös Európai Referenciakeret (KER) szintleírásaiban megadott szintekhez való hozzáigazítása. Ennek kapcsán nyilvánvalónak tűnik, hogy a KER által definiált közös hivatkozási keretben a szintekhez rendelt megfelelési minimumokat is lehetséges volna standardizálni, mivel a logit értékeket a horgonyfeladatok segítségével történő, egymástól egyébként akár lényegesen eltérő tesztek összekapcsolása által szintén közvetlenül összehasonlíthatóvá tehetjük. Ez azért lehetséges, mert az IRT-alapú elemzések segítségével az összekapcsolt tesztek-

ben szereplő itemek nehézségi indexét – s ezáltal a vizsgázók képességindexét is – közös referenciaponthoz tudjuk viszonyítani.

Ez a közvetlen összehasonlíthatóság más területeken is hasznos lehet. Mivel az egyes vizsgázók teljesítménye azonos viszonyítási ponthoz kötött és mivel a vizsgázók képességindexe közötti különbség intervallum skálán követhető, lehetséges például egy adott vizsga különböző verzióinak kapcsán a vizsgázói populáció összetételének, illetve az összetétel esetleges változásainak nyomon követése és beható elemzése.

További előnyt jelent az a tény, hogy amennyiben a megfelelési minimumokat logit pontokban határozzuk meg, nincs szükség annak garantálására, hogy egy adott vizsga különböző verziói teljes mértékben ekvivalens tesztváltozatokat alkalmaznak. Mivel a tesztverziók egymáshoz vannak kapcsolva, a tesztitemek nehézségi szintje a korábbi tesztverziókban alkalmazott itemek nehézségi szintjéhez képest kerül meghatározásra. Ennek függvényében állapítható meg a vizsgázók képességindexe, s ezért lehet fix a megfelelési minimum ily módon definiált szintje.

A számos előny mellett azonban, mint azt jeleztük, ki kell térnünk a potenciális hátrányokra is. Ezek közül talán a leglényegesebbnek a képesség logit pontok alkalmazásának lehetséges felszíni validitás problémái tűnnek.

A felszíni validitás kérdésében a szakirodalom némiképp megosztott, amennyiben egyes források nem tekintik olyan súlyú tényezőnek, mint a validitás egyéb aspektusait. Mindazonáltal joggal tételezhetjük fel, hogy amennyiben a teszttel kapcsolatba kerülő „laikusok" nagy számban érzik úgy, hogy a teszt érvényessége aggályos, akkor ezt a tényt aligha lehet figyelmen kívül hagyni. Márpedig – pontosan a képességindexek ponthatárok helyett történő alkalmazásának kevéssé elterjedt mivolta okán – ez az eljárás vélhetően sokak számára kétes érvényűnek tűnhet.

Képzeljük el például, hogy két azonos pontszámot elérő vizsgázó közül az egyik megfelel a vizsgán, a másik azonban – eltérő képesség logit pontja miatt – nem. Vajon elfogadja-e a tesztelméletben kevéssé jártas vizsgázó a modern tesztelméleten alapuló, számára részleteiben aligha követhető magyarázatot? Vajon a ponthatárok világában szocializálódott vizsgázók el tudják-e fogadni ezt a fajta paradigmaváltást, amely azt jelenti: többé nem a laikus számára is átlátható és könnyen érthető pontszámok, hanem – valamiféle szakmai elit által meghatározott és bonyolult számításokba bugyolált – „képességindexek" határozzák meg, kinek volt sikeres a vizsgája? És vajon vállalhatóak-e oktatáspolitikai szempontból azok a konfliktusok, amelyek mindebből következhetnek?

Pillanatnyilag nehéz a fenti kérdésekre egyértelmű válaszokat adni, bár a probléma felvetése relevánsnak tűnik. Hiba volna azonban az felszíni validitás esetleges problémái miatt azonnal elvetni a képességindexek alkalmazását. Éppen az eljárás viszonylagos ismeretlensége a nehézségek fő forrása, vagyis a vizsgázói közvélemény megfelelő tájékoztatása és a vizsgázók számára is megfogható előnyök bemutatása orvosolhatja a problémát.

A gyakorlati alkalmazás szempontjából ugyancsak potenciális nehézséget jelenthet az a tény, hogy az IRT-alapú statisztikai számítások csak viszonylag nagy vizsgázói populációk esetében adhatnak megfelelően megbízható eredményt. Ez azt jelenti, hogy a több száz vagy több ezer fős vizsgák esetében ugyan nincs probléma, de a száz fő alatti tesztek és vizsgák esetében a modern tesztelméleti elemzések nem használha-

tóak, s a hazai nyelvoktatási kontextusban nem ritkák az ilyen viszonylag alacsony létszámok.

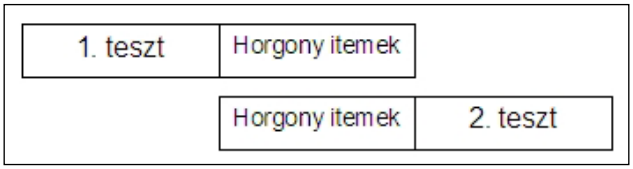
Erre a problémára megoldást jelenthet a feladatbankok alkalmazása. Bár feladatbanknak nevezett feladatgyűjtemények ma is nagy számban léteznek, valódi feladatbankoknak csak az olyan gyűjtemények tekinthetők, amelyekben minden egyes item pszichometriai jellemzői pontosan meg vannak határozva. Ez azt jelenti, hogy például az itemek nehézségi szintje ismert, kalibrált érték, s ezen túl más olyan kvantifikálható jellegzetességek is (például illeszkedés) dokumentáltak, amelyek világossá teszik az itemek felhasználhatósági körét. Ha tehát rendelkezünk ilyen kalibrált nehézségű feladatokkal, akkor az ismert nehézségű itemekből ismert nehézségű tesztek állíthatók elő, s ily módon akár egyetlen vizsgázó képességszintje is megállapítható az itemekre adott válaszok tükrében.

Mi több, a feladatbankok segítségével olyan tesztek is készíthetők, amelyeknek a nehézségi szintje alkalmazkodik a vizsgázók képességszintjéhez. Az ilyen adaptív teszteket számítógépes felületen oldják meg a vizsgázók. Minden válasz után a számítógép, a válasz helyességétől függően, új, nehezebb vagy könnyebb itemet kínál fel a vizsgázónak. Megfelelő számú item megválaszolása után a képességindex értéke már könnyen meghatározható. Ily módon nem csak az garantálható, hogy a vizsgázók képességét pontosabban lehet meghatározni (hiszen a teszt a vizsgázó tudásszintjéhez idomul), hanem azt is, hogy gyakorlatilag minden vizsgázó más-más tesztet old meg, ami vizsgabiztonsági szempontból is igen előnyös (Baker 1997 :50).

Mindehhez azonban olyan feladatbankokra van szükség, amelyek nagy számú kalibrált itemet tartalmaznak, s az ilyen feladatbankok létrehozása igen költséges és időigényes vállalkozás. Erre tett kísérletet a nemzetközileg is nagy hírű holland nemzeti mérési központ (CITO) által gondozott európai idegen nyelvi feladatbank projekt (EBAFLS). Az Európai Bizottság által szponzorált vállalkozás célja olyan feladatbank létrehozása volt, amelyben a KER B1-es szintjéhez kapcsolt olvasásértési és hallásértési feladatok találhatóak három nyelven (angol, német, francia). A projekt 2007 szeptemberében lezárult, ám az eredmények elmaradtak a várakozásoktól, elsősorban azért, mert a projekt tapasztalatai szerint a KER szintjeinek item nehézségi szintek megállapításához történő alkalmazása kapcsán nem sikerült konszenzust kialakítani a részt vevő országok szakértői között (EBAFLS 2008). A projekt részleges kudarca azonban nem magát a feladatbank-építés alapjait kérdőjelezi meg. Mindezt igazolni látszik számos sikeres feladatbank-építési projekt (például Henning 1986; Leclercq és Bruno 1993; Szabó 2008).

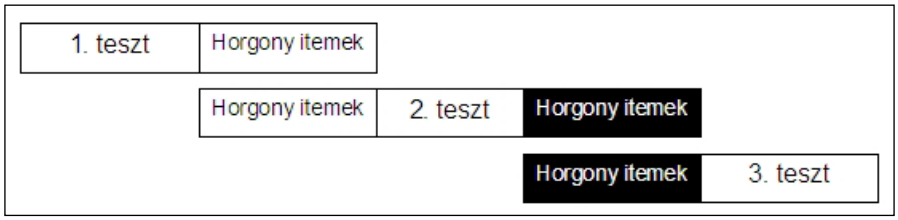
A feladatbankok előállítása és gyakorlati alkalmazása azonban még optimális feltételek között is meglehetősen időigényes feladat, s a feltételek sokszor távolról sem optimálisak. Valódi lehetőség azonban a különböző tesztverziók korábban már említett horgony itemek segítségével történő összekapcsolása, ami lényegesen egyszerűbben és gyorsabban kivitelezhető.

Természetesen ennek az eljárásnak is akadnak hátrányai. Elsősorban azt kell ezen a ponton kiemelni, hogy a különböző tesztverziókat alapesetben *közös* itemek kötik össze (4. ábra), vagyis ha egy vizsgázó ismeri az első teszt itemeit, a második teszt esetében ezek az itemek már ismertek lesznek számára, s ez nyilván befolyásolhatja a teljesítményét.




***4. ábra.*** **Tesztek összekapcsolása közös itemekkel (1)**

A közvetlen összekapcsolás azonban nem az egyetlen lehetséges megoldás. Mint az 5. ábrán látható, a különböző tesztverziók nem feltétlenül ugyanazokkal az itemekkel kell hogy összekapcsolódjanak. Az 5. ábrán szereplő 1. teszt és 3. teszt közvetlenül már nem kapcsolódik össze, közös itemeket nem tartalmaz. Ily módon akár ugyanabban a populációban is problémamentesen felhasználható.




***5. ábra.*** **Tesztek összekapcsolása közös itemekkel (2)**

Mindezek alapján joggal állíthatjuk, hogy a tesztverziók összekapcsolása praktikusan is lehetséges, bár kétségkívül nem egyszerű feladat. Amennyiben azonban sor kerül az összekapcsolásra, nem csupán a megfelelési minimumok stabilitása növelhető, de arra is mód nyílik, hogy a vizsgázói populáció képességszint-összetételének esetleges változásairól is képet kaphassunk. Nagy vizsgázói populációk esetében vélhetően kisebb az esélyük – bár nem zárhatók ki – a vizsgázók képességszintjét érintő drámai változásoknak, ám a kisebb, legfeljebb néhány száz fős vizsgák esetében komolyan kell számolni azzal a lehetőséggel, hogy a populáció összetétele vizsgánként akár szignifikánsan is átalakulhat. Ez a tény pedig különleges súllyal esik latba a ponthatárok megállapítása szempontjából, hiszen amennyiben a teljesítmény-minimumokat a korábban vázolt módon a kritérium-orientált és norma-orientált eljárások kombinálásával állapítják meg, a populáció képességszint-változásai esetenként lényegesen is kihathatnak a ponthatárra. Mindez azonban nem következhet be, ha a megfelelési minimum képességszint index formájában kerül meghatározásra.

## 5. Következtetések

Mint a fentiekből kiderül, az IRT-alapú képességindexek ígéretes lehetőséget jelentenek a teljesítményminimumok hatékonyabb meghatározásához. A mintafüggetlen elemzési eljárások során megállapított vizsgázói képességindexek objektívebbek a nyerspontoknál; használatuk segítségével pontosabban, a mérés céljainak jobban megfelelő módon dönthető el, hogy egy adott vizsgázó teljesítménye elegendő-e a minimális megfeleléshez. Az IRT elemzések technikai háttere adott, számítógép alapú alkalmazásuk mind hardver, mind szoftver szempontjából problémamentes.

Tagadhatatlan ugyanakkor, hogy a gyakorlatban a képességindex-alapú teljesítményminimum-meghatározás számos esetben nehézségekbe ütközik. Ezek egy része olyan objektív okra vezethető vissza, amely akár tartósan is megakadályozhatja a képességindexek alkalmazását. Leghangsúlyosabb problémaként a vizsgázók létszámát kell itt kiemelnünk, bár a feladatbankok alkalmazása képes megoldani a kis vizsgázói létszámok miatt előálló nehézségeket is. Tény azonban, hogy nagy számú, széles körben felhasználható feladatbank előállítására jelenleg kevés esély mutatkozik, így a kis vizsgázói létszámok esetében (például osztálytermi mérések kapcsán) a fenti eljárás kevéssé hasznosítható. Fontos azonban azt is hangsúlyozni, hogy a feladatbankok létrehozása ugyanakkor reális és kívánatos cél a nagy vizsgázói létszámokkal operáló vizsgák esetében, hiszen mint fentebb láthattuk, a feladatbankok nagyban hozzájárulhatnak a mérés színvonalának emeléséhez. Mi több, a létrejövő feladatbankok elméletileg különféle populációk számára is felhasználhatóak lehetnek, azaz végső soron még a kis létszámú vizsgák is profitálhatnak belőlük. Ehhez azonban magas szintű, akár nemzetközi kooperációra van szükség, s bár az EBAFLS részleges sikertelensége óvatosságra int, a lehetőséget hiba volna elvben is elvetni.

Lényegesebbnek tűnik azonban az olyan gátló tényezők megszüntetése, amelyek nem elméleti alapon jelentenek nehézséget, hanem csak az eljárás ismeretlensége miatt okoznak problémát. Ennek kapcsán joggal bízhatunk abban, hogy amennyiben a tágabb értelemben vett szakmai közösség, illetve maguk a vizsgázók felismerik a képességindexek alkalmazásában rejlő lehetőségeket, nem lesz majd akadálya az ilyen jellegű eljárások szélesebb körben történő bevezetésének. Bár, mint fentebb jeleztük, jelenleg a képességindexek közvetlen alkalmazása vélhetően inkább bizonytalanságot és bizalmatlanságot eredményezne, végső soron az IRT-alapú eljárás éppen úgy a vizsgázó érdekeit is szolgálja, mint a vizsgarendszerekét. Ha a megfelelési minimumok képességindex-alapú meghatározása érvényesebb mérést eredményez, a vizsgázók is nagyobb bizalommal fordulhatnak az ilyen típusú eljárások felé. Ennek pedig megnövekedett bizalom és megnövekedett felszíni validitás lesz az eredménye.

A jövő útja tehát vélhetően a képességindex-alapú megfelelési minimumoké, hiszen a potenciális előnyök hosszú távon mindenképpen felülmúlják a pillanatnyi hátrányokat. Nagy kérdés azonban, hogy a sikerhez szükséges változások lehetősége mikorra érik tényleges változássá, hiszen az élet szinte valamennyi területén igaz az a tétel, miszerint a „status quo" megőrzése mindig a legkényelmesebb megoldás, s ez nincs másként a nyelvtudásmérés esetében sem.

Biztató jel azonban, hogy az államilag elismert nyelvvizsgák közül egyre több alkalmaz a klasszikus elemzési metódusokon túl modern tesztelméleti megközelítést is, s ez a tény növeli a változás esélyeit. Hadd reméljük azt is, hogy a nyelvtudásmérés – és egyben a pedagógiai mérések – egyik leglényegesebb hazai színterén, az érettségi vizsga keretein belül is mielőbb teret nyernek a korszerű elemzési és értékelési eljárások. Ha pedig az IRT-alapú elemzések elterjednek, hamarosan a képességindexek szerepe is felértékelődhet, elősegítve a teljesítményminimum-meghatározás modern módszereinek elterjedését.

A pedagógiai mérések s így a nyelvtudásmérés kapcsán is gyakran emlegetett dilemma, vajon mennyire lehet egyáltalán éles határt vonni a még éppen megfelelő és a már nem megfelelő teljesítmények között, különös tekintettel arra, hogy magának

a nyelvtudásnak mint konstruktumnak a meghatározása is máig meglehetősen homályos (vö. például Bárdos 2002). Nyilvánvaló, hogy a képességindexek alkalmazása sem adhat minden részletében kielégítő választ erre a kérdésre. Mindemellett a megfelelési minimumok lehető legszakszerűbb meghatározása hozzásegítheti mind a mérési szakembereket, mind a vizsgaeredmények felhasználóit, köztük magukat a vizsgázókat is, hogy a teljesítményeket megfelelően értelmezzék. Ez pedig nem jelent mást, mint hogy minden mérés legfőbb célja, az érvényesség is javulhat, közelebb hozva a mérést mint mesterséges közeget a nyelvhasználat autentikus közegéhez.

## IRODALOM


Alderson, J. C. – C. Clapham – D. Wall (1995): *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.

Angoff, W. H. (1971): Scales, norms and equivalent scores. In: Thorndike, R. L. (szerk.): *Educational Measurement*. 2nd ed. Washington, D.C.: American Council on Education, pp. 508–600.

Bachman, L. F. (2004): *Statistical Analyses for Language Assessment*. Cambridge: Cambridge University Press.

Baker, R. (1997): *Classical Test Theory and Item Response Theory in Test Analysis*. Special Report No 2: Language Testing Update.

Bárdos Jenő (2002): *Az idegen nyelvi mérés és értékelés elmélete és gyakorlata*. Budapest: Nemzeti Tankönyvkiadó.

*Building a European Bank of Anchor Items for Foreign Language Skills* (*EBAFLS*) (2008): http://www.cito.com/research_and_development/participation_international_research/ebafls.aspx Letöltve: 2009. május 8.

Henning, G. (1986): Item banking via DBASE II: the UCLA ESL proficiency examination experience. In: Stansfield, C.W. (szerk.): *Technology and Language Testing*. Washington, D.C.: TESOL.

Hambleton, R. K. – B. S. Plake (1995): Extended Angoff procedures to set standards on complex performance assessments. *Applied Measurement in Education* 8, pp. 41–56.

Hambleton, R. K. – H. Swaminathan (1985): *Item Response Theory*. Boston: Kluwer-Nijhoff.

Hambleton, R. K. – H. Swaminathan – H. J. Rogers (1991): *Fundamentals of Item Response Theory*. Newbury Park: SAGE Publications.

Hulin, C. L. – F. Drasgow – C. K. Parsons (1983): *Item Response Theory*. Homewood, Illinois: Dow-Jones Irwin.

*Közös európai referenciakeret: nyelvtanulás, nyelvtanítás, értékelés* (2002): Pilisborosjenő: Pedagógusképzési Módszertani és Információs Központ.

Leclercq, D. A. – J. E. Bruno (1993, szerk.): *Item Banking: Interactive Testing and Self-Assessment*. Berlin; Heidelberg: Springer-Verlag.

Livingston, S. A. – M. J. Zieky (1982): *Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests*. Princeton, New Jersey: Educational Testing Service.

McNamara, T. F. (1996): *Measuring Second Language Performance*. London; New York: Longman.

— (2000): *Language Testing*. Oxford: Oxford University Press.

*Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment* (*CEFR*). *A Manual*. (2009). Strasbourg: Language Policy Division, Council of Europe.

Szabó, Gábor (2008): *Applying Item Response Theory in Language Test Item Bank Building*. Frankfurt am Main: Peter Lang.