MAGYAR PEDAGÓGIA
108. évf. 1. szám 29–51. (2008)

EGY IRT-ALAPÚ NYELVI FELADATBANK LÉTREHOZÁSÁNAK MÓDSZERTANI KÉRDÉSEI
A német érettségi vizsgafeladatok elemzésének eredményei

Vígh Tibor
Szegedi Tudományegyetem, Neveléstudományi Doktori Iskola

A nyelvoktatás és a nyelvtudásmérés területén az elmúlt évtized legfontosabb és legnagyobb hatású dokumentuma a *Közös európai referenciakeret: nyelvtanítás, nyelvtanulás, értékelés* (2002)[1], amely „közös alapot teremt Európa-szerte a nyelvi tantervek, tantervkészítési irányelvek, vizsgák, tankönyvek stb. kidolgozásához" (*KER*, 2002. 1. o.). A dokumentum egységes alapot nyújt a kommunikatív nyelvtudás értelmezéséhez, mivel a Referenciakeret a korábbi nyelvtudás-modellekre építve pontosan definiálja a kommunikatív nyelvi kompetenciák fogalmát, alkotóelemeit és dimenzióit. A dokumentum továbbá egy fokozatosan emelkedő, egymásra épülő referenciaszint-rendszert vázol, amely alapján lehetővé válik, „hogy a tanulók haladása a nyelvtanulás minden szakaszában és az egész életük során mérhető legyen" (*KER*, 2002. 1. o.), hozzájárulva ahhoz, hogy a nyelvtudást mérő vizsgák szintjei összehasonlíthatók legyenek (*Bárdos*, 2006).

Ennek a folyamatnak köszönhetően a nyelvtudásmérésben egy újabb kutatási irány bontakozott ki, amelynek célja a különböző nyelvtudást mérő vizsgák szintjeinek validálása a Referenciakerethez. E folyamat alapjául szolgál a *Nyelvvizsgák szintillesztése a Közös európai referenciakerethez* (*North*, *Avermaet*, *Figueras*, *Takala* és *Verhelst*, 2003) kézikönyvben leírt szintillesztő eljárás és a hozzá tartozó módszertani segédlet (*Banerjee*, *Kaftandjieva*, *Takala* és *Verhelst*, 2004). Európai projektek is elindultak azzal a céllal, hogy a Referenciakeret szintleírásai alapján nyelvi feladatbankot hozzanak létre. Ide tartozik például a DIALANG-projekt (*Diagnostic Language Assessment System for Learners*) és a 2008-ban lezárult EBAFLS-projekt (*Building a European Bank of Anchor Items for Foreign Language Skills*). Ebben a kutatási irányban – bár vannak fontos eredmények – a különböző nyelvtudást mérő vizsgák szintjeinek illesztése a Referenciakerethez még nem történt meg.

A hazai nyelvi mérés és értékelés területén a különböző nyelvvizsgák esetén is fontos kutatási feladat a nyelvi szintek kialakítása és kalibrálása, továbbá nyelvi feladatbank létrehozása. A magyar közoktatási rendszerben a 2005-ben bevezetett idegen nyelvi érettségi jelentős fejlődésen ment keresztül, hiszen célként jelenik meg a kommunikatív nyelvtudás mérése és értékelése, valamint a Referenciakeret szintleírásához történő il-

[1] A tanulmányban a Referenciakeret fogalmát használjuk, ha a dokumentumról általában van szó. A rövidített hivatkozás: *KER*, 2002.

leszkedés. Ugyanakkor jelenleg még nem áll rendelkezésünkre olyan kutatási eredmény, amellyel az idegen nyelvi érettségi két szintjének egymáshoz és a Referenciakeret szintrendszeréhez történő illeszkedését empirikusan igazolhatnánk.

A nyelvtudás mérésében, a különböző nyelvi szintek kialakításában, valamint a nyelvi feladatbank kiépítésében a kommunikatív tesztelés támaszkodik a valószínűségi (modern) tesztelmélet (*Item Response Theory*, *IRT*) eszközeire. A nyelvtudásmérésben a valószínűségi tesztelméleti modelleket a nyolcvanas évek közepe óta használják (*Nikolov*, *Pércsich* és *Szabó*, 2000). Az elmúlt 10-15 évben jelentős szerephez jutott a nyelvvizsgák feladatainak megírásában, elemzésében és kiértékelésében (*Eckes*, 2003), valamint a nyelvtudást mérő vizsgáknak a Referenciakerethez történő illesztésében vált domináns eszközzé.

Tanulmányunkban a német érettségi vizsgafeladatok nehézségi szintjének vizsgálata alapján mutatjuk be, hogyan építhető ki egy, a Referenciakeret szintrendszerére épülő IRT-alapú nyelvi feladatbank, valamint e folyamat során milyen módszertani kérdések, problémák merülhetnek fel. E cél eléréséhez több szempontból közelítjük meg a témát. Először a valószínűségi tesztelmélet alapjait, modelljeit és a nyelvtudásmérésben való alkalmazhatóságának lehetőségeit mutatjuk be. Ezt követően térünk ki a nyelvi szintek empirikus kialakításának szükségességére és problémáira az idegen nyelvi érettségin. A harmadik részben egy adatbázis-elemezés eredményeit mutatjuk be, amelynek során a 2006. évi középszintű német nyelvi érettségi olvasott és hallott szöveg értése és a nyelvhelyesség vizsgafeladatok eredményeit vizsgáltuk a dichotóm adatok elemzésére és értelmezésére alkalmas Rasch-modellel. Végül kitérünk arra is, hogyan és milyen európai projektek eredményeit felhasználva lehetne a valószínűségi tesztelméleti eszközök alkalmazásával az idegen nyelvi érettségi vizsga szintjeinek egymáshoz és a Referenciakerethez történő illeszkedését vizsgálni.

A kommunikatív nyelvtudás mérése és az IRT-modellek

A Referenciakeretben definiált kommunikatív nyelvi kompetenciák fogalma egy többkomponensű, közvetlenül nem mérhető és értékelhető látens konstruktumot (*latent trait*) jelöl. E komplex képesség szintjeinek és összetevőinek mérése és értékelése indirekt módon, a receptív (beszédértés, olvasás) és a produktív (beszéd és írás) készségek alapján történik (*Alderson*, *Clapham* és *Wall*, 1995). Ugyanakkor a nyelvtudást mérő vizsgák keretében történő nyelvi készségek megítélésének alapvető korlátai vannak, számolni kell azzal, hogy a tesztfeladatban megadott nyelvi szituációk (az autentikusságra való törekvés ellenére) mesterségesek, a mérés nem kívánt beavatkozás; a vizsgahelyzet irreális és nyelvezete természetellenes (*Bárdos*, 2003).

Ezen problémák enyhítésére a kommunikatív tesztelés kiemeli a nyelvi tesztek validitásának fontosságát, valamint újabb típusait, aspektusait is meghatározza. Ezek közé tartozik az autentikusság, az interaktivitás és a teszthatások, amelyeknek a biztosítása a

teszt érvényességének és megbízhatóságának az előfeltételeként fogható fel[2] (*Bachman* és *Palmer*, 1996). A kommunikatív tesztelés a nyelvi készségek mérése és értékelése során olyan teszteket alkalmaz, amelyek a látens képességet mozgósítják a tesztfeladatok és az egyes tesztelemek, az itemek segítségével (*Bárdos*, 2002). A tesztfeladatokra adott reakciók alapján így egyrészt lehetővé válik, hogy a vizsgázók nyelvtudását mérjük (*Grotjahn*, 2000), másrészt pedig, hogy a teszten elért eredmények alapján megfelelő döntéseket hozzunk.

A valószínűségi tesztelméleti eszközök alkalmazásának is az az alapfelvetése, hogy egy tesztitemre adott helyes vagy helytelen választ fizikailag nem megfigyelhető olyan látens tulajdonság magyarázza (*Molnár*, 2006), amely a feladatokat megoldó személy válaszainak konzisztenciájáért felelős (*Wainer* és *Messick*, 1983. 343. o. idézi *Nikolov*, *Pércsich* és *Szabó*, 2000. 10. o.). Ez a látens tulajdonság azt is meghatározza, hogy egy személy az adott itemet milyen valószínűséggel oldja meg (*Eckes*, 2003).

Az IRT-modellek egyszerre követik nyomon, hogy a vizsgázó milyen választ adott és milyen válasz lett volna a legvalószínűbb. Az így kapott értékeket a modellek folyamatosan összevetik, meghatározzák az eltéréseket és így eljutnak a szisztematikus tényezők mintafüggetlen becsléséig (*Dávid*, 2006) és a képességparaméterek tesztfüggetlen meghatározásáig (*Molnár*, 2008). „Mindegyik IRT-modellben közös, hogy (1) adott item esetén megadják a személy helyes válaszadásának valószínűségét, (2) nem determinisztikusak, hanem valószínűségi alapokon nyugszanak, illetve (3) 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" (*Molnár*, 2006. 101. o.). Ezeket a modelleket csoportosíthatjuk az összefüggések típusa és az itemparaméterek száma szerint (*Molnár*, 2003). A továbbiakban három modell alkalmazási lehetőségét mutatjuk be a nyelvtudásmérésben.

A Rasch-modell (*Rasch's simple logistic model*) dichotóm adatok elemzésére alkalmas. A modell az item nehézségét és a vizsgált személy képességét kezeli egy paraméterként és az itemek diszkriminációs indexét azonosnak veszi (*Molnár*, 2006). A nyelvi mérés és értékelésben főként a receptív készségek dichotóm itemeinek vizsgálatára alkalmas. A modell főbb tulajdonságaira, a Rasch-modellel kapott eredmények értelmezésére a tanulmány további részében térünk ki.

A nyelvtudást mérő nemzetközi és hazai vizsgák produktív készségeket mérő vizsgarészében az egységesebb és objektívebb értékelés biztosítása érdekében elterjedtek az analitikus értékelési skálák, amelyek kritériumorientált értékelést tesznek lehetővé. Ennek az értékelési eljárásnak a lényege, hogy a maximális pontszámot leíró deskriptor az elérendő kritériumot reprezentálja, amelynek további felbontásával határozzák meg a többi skálafokot, amelybe a vizsgált személy teljesítménye besorolható. Az értékelési skálák alkalmazásával a vizsgált személy képességszintjét az elvárt szinthez viszonyítják, így a skálák működése a kritériumhoz mint viszonyítási ponthoz való megfeleltetés alapján történik (*Vígh*, 2007b). Gyakran előfordul, hogy egy skálafokhoz két (vagy több) pontszám tartozik, így a pontszámok nem azonos lépésközre találhatók egymástól, az

[2] A kommunikatív tesztek jóságmutatóit *Bachman* és *Palmer* (1996) alapján magyar nyelven *Vígh* (2005, 2007a) összegzi.

értékelési skáláknak tehát gyakran eltérő a skálaszerkezete. Ebben az esetben alkalmazható *Masters* parciális kredit modellje (*partial credit model*), amellyel azt lehet minősíteni, hogy a vizsgáztatók hogyan alkalmazzák az értékelési skálákat, valamint hogy minden deskriptort egyenlő arányban használnak-e (*Dávid*, 2006). A modell további előnye, hogy lehetővé teszi a dichotóm és nem dichotóm itemek közös elemzését is, mivel a parciális kredit modell a Rasch-modell kiterjesztett változata (*Molnár*, 2008).

A produktív készségek értékelése során gyakran találkozunk értékelési hibával, különböző értékelők esetenként más értékelési sávba sorolják ugyanazon vizsgált személy teljesítményét. Az eredményt ebben az esetben a személy képességei, a feladat és az értékelési szempontok nehézsége, valamint az értékelő szigorúsága befolyásolja (*Eckes*, 2003, 2004; *Molnár*, 2003). Ezeket a tényezőket veszi figyelembe *Linacre* sokoldalú modellje (*multifaceted model*). A nyelvtudást mérő vizsgákon (például *TestDaF*, *Eckes*, 2003, 2004) a produktív készségek értékelésekor ezt a modellt alkalmazzák az igen költséges kettős értékelés helyett, mivel a modell alkalmazásával megbízhatóbb eredményeket kapunk. A kettős értékelés során csak két értékelő döntése közötti (nem)egyezést lehet vizsgálni, és nem tudjuk megállapítani, hogy mennyire felelnek meg az értékelők a standard értékelésnek. A sokoldalú modell azonban lehetővé teszi, hogy az értékelők csoportjában határozzuk meg az értékelők szigorúságát, így a vizsgázók produktív készségeinek mérése megbízhatóbb lesz.

A valószínűségi tesztelméleti modellek az idegen nyelvi mérés és értékelés területén széles körben alkalmazhatók, mivel sok a hasonlóság a nyelvtudás mérésének lehetőségei és a valószínűségi tesztelméleti eszközök alapfelvetése között. A nyelvtudást mérő vizsgák szintjeinek kialakításában, egységesítésében, a nyelvi feladatok nehézségi szintjének mérésében, valamint az értékelési eljárások megbízhatóságának növelésében az IRT-modellek fontos szerepet töltenek be. A hazai nyelvtudásmérésben a valószínűségi tesztelméleti modelleket egyre gyakrabban alkalmazzák, az egyes nyelvi feladatok nehézségének vizsgálatában több hazai empirikus kutatás eredményeit is ismerjük (pl. *Alderson*, 2000; *Dávid*, 2005, 2006; *Nikolov*, *Pércsich* és *Szabó*, 2000).

A nyelvi szintek problematikája az idegen nyelvi érettségin

Az 1980-as években egyre inkább előtérbe kerültek a közoktatásban megszerezhető nyelvtudással szemben támasztott minőségi elvárások. A kilencvenes évek elejétől megjelent az a társadalmi igény, hogy a diákok azonnal, közvetlenül hasznosítható, és a munkaerőpiacon, illetve a felsőoktatásban helyzeti előnyt jelentő nyelvtudást szerezzenek. Az idegen nyelvek tudásának presztízse nagymértékben felértékelődött.

A társadalom és a civil szféra elvárásainak megfelelően a *Nemzeti alaptanterv* (1995) és a 2000-ben megjelent *Kerettantervek* deklarált célja a mindennapi életben használható nyelvtudás elsajátíttatása a közoktatásban. A használható, kommunikatív nyelvtudás a *Nemzeti alaptanterv* (NAT) 2003-as módosítása szerint „az adott [nyelvi] szituációnak megfelelő nyelvhasználati képességet jelenti" (*NAT*, 2003. 39. o.). Az alaptanterv (*NAT*, 2003) iskolai követelménnyé teszi a Referenciakeret egyes szintjeinek elérését. A közok-

tatásban elérendő szint az első idegen nyelv esetén a 12. évfolyam végére a B1-B2, míg a második idegen nyelv esetén az A2-B1 szint. A Referenciakeret hatfokú skáláján található „mesterfokú nyelvhasználói szintek [C1 és C2] elérése a közoktatásban nem tekinthető alapfeladatnak" (*NAT*, 2003. 40. o.).

E változásoknak köszönhetően az idegen nyelvi mérés és értékelés területén jelentős fejlődésként értelmezhető az új idegen nyelvi érettségi vizsga, amelynek bevezetését az alaptantervben (*NAT*, 1995, 2003) megjelenő kommunikatív nyelvi kompetencia és szintjeinek mérése tette szükségessé. A korábbi érettségi vizsga nem követte a kommunikatív szemléletet, elsősorban a nyelvtani-lexikai kompetenciát mérte, nem alkalmazott például autentikus szövegeket, és a hallott szöveg értésének méréséről lemondott (*Petneki*, 2007). A vizsgareform során (lásd *Alderson*, *Nagy* és *Öveges*, 2000; *Einhorn*, 2004) az idegen nyelvi érettségit úgy alakították ki, hogy mérési és értékelési eljárásaiban kövesse a kommunikatív tesztelés alapelveit, valamint illeszkedjék a Referenciakeret szintleírásaihoz. A *Nemzeti alaptanterv*ben (2003) deklarált céloknak megfelelően az idegen nyelvi érettségi vizsga két szintjével (közép- és emelt szint) a Referenciakeret három szintjét, az A2-B1 és a B2 szintet méri. A középszintű érettségi vizsgán az A2-B1 szint mérésére szükség van, mivel az érettségi egy rendkívül heterogén közoktatási rendszer záróvizsgája, ahol a különböző tanulói teljesítmények között nagyon jelentős különbségek vannak, amelyekhez a feladatsoroknak illeszkedni kell (*Einhorn*, 2006). A vizsgázók nyelvi képességeinek minél teljesebb mérése érdekében minden feladatsor a lépcsőzetesség elve alapján épül fel, a feladatok és az itemek egyre nehezebbek (*Einhorn*, 2007b). Komoly kihívást jelent a különböző vizsgaidőszakokban keletkező feladatsorok nehézségi szintjének, állandóságának biztosítása. Az emelt szint a Referenciakeret B2 szintjét méri, így mérésmetodikai szempontból megfelelőbb. A hazai nyelvtudásvizsgálatok eredményei azonban azt mutatják, hogy a diákok a C1 szintet az iskolarendszeren belül a nyelvi tagozatokon, a két tanítási nyelvű iskolákban el tudják érni, így a jelenlegi vizsgaszint nem jelent számukra kihívást.

Az idegen nyelvi érettségi vizsgának jelentős szerepe lehet a közoktatásban megszerezhető nyelvtudásszint monitorozásában, változásának nyomon követésében, hiszen jelenleg ez még megoldatlan (*Nikolov*, 2007). A magyar érettségizők nyelvtudásszintjéről, az idegen nyelvi érettségi feladatsorok nehézségi szintjéről jelenleg kevés adat áll rendelkezésre (*Einhorn*, 2006). Ezen a területen számos fejlesztő kutatásra van szükség. Egyrészt szükséges a vizsgázók reális nyelvtudásszintjének vizsgálata annak érdekében, hogy ahhoz tudják igazítani a vizsgát minden tanított idegen nyelvből. Másrészt vizsgálni kell, hogy a vizsgafeladatok szintje hogyan viszonyul a követelményekben rögzített készségszintekhez (*Einhorn*, 2007b).

Vígh Tibor

A német nyelvi érettségi feladatok adatbázisának elemzése

Az elemzés előzményei, célja

Az elemzés során a 2005. és 2006. évi német nyelvi érettségi közép- és emelt szintű vizsgafeladatokhoz tartozó adatbázisokat alkalmaztuk[3]. A két év német nyelvi érettségi feladatsor előkészítéséről, azok elemzéséhez összeállított adatbázisokról, az elemzés módszereiről és eredményeiről lásd *Einhorn Ágnes* (2007a, 2008) tanulmányait. Mindkét évben a közép- és emelt szintű adatbázis az olvasott és hallott szöveg értése, a nyelvhelyesség feladatsorok itemszintű adatait és az íráskészség vizsgarész részpontszámait tartalmazza. A négy adatbázist csak külön elemezhetjük, hiszen nincs olyan item vagy személy, amely összekötné azokat, ehhez újabb felmérésre lenne szükség. Ebből adódóan csak a 2006. évi középszintű érettségi[4] adatbázisában szereplő két receptív készség, valamint a nyelvhelyesség Rasch-modellel történő elemzését mutatjuk be.

Az adatbázis a teljes populációból (N=26239) 1092 vizsgázó adatát tartalmazza. Ezekből az adatokból nem lehet automatikusan következtetni a vizsgázók teljesítményére, de az egyes feladatok empirikus nehézségének vizsgálatához elegendő információt nyújtanak. Az adatbázis önálló vizsgálatát indokolja az a tény, hogy a vizsga célja a Referenciakeret két szintjének (A2-B1) mérése, így a vizsgafeladatoknak egy nagyon széles képességterületet kell átfogni. A vizsgafeladatok előzetes kipróbálása, nehézségi szintjük vizsgálata nem történt meg, ezért ezt az adatbázist az alapján elemezzük, hogy a feladatsorok empirikus nehézségi szintje mennyre fedi le a vizsgázók képességtartományát.

A 2006. évi német érettségi eredményeinek elemzése során (*Einhorn*, 2008) vizsgálták a feladatsorok teljesítettségét, minőségét, a feladatsorok megoldásakor megjelenő tipikus tanulói problémákat, valamint a feladatok értékelésének minőségére vonatkozó információkat. A receptív készségeket és a nyelvhelyességet mérő füzetekben vizsgálták az egyes feladatok itemeinek működését a különböző teljesítménycsoportokban. Az elemzés során az itemek nehézségét az összpontszámhoz viszonyítva adták meg, amely alapján meghatározható, hogy a mintában és a teljesítménycsoportokban milyen arányban válaszolták meg helyesen a vizsgázók az adott itemet. Ez az elemzés lehetővé teszi a kiugróan könnyű és nehéz itemek azonosítását, de az nem állapítható meg, hogy az itemek nehézségi indexei között mekkora különbségek vannak. Egy item nehézségi szintje a klasszikus tesztelméleti módszerekkel számolva „nemcsak az itemet, hanem a mintát is jellemzi [és] a tanulók képességének fejlettségéről kapott kép függ a teszt jellemzőitől is" (*Molnár* és *Józsa*, 2006. 158. o.). Az adott feladatsor nehézsége főként az átlagos képességű vizsgázók eredményét befolyásolja (*Molnár*, 2006). A nem megfele-

[3] A 2005. évi középszintű német nyelvi érettségi vizsga adatbázisa 1144, a 2005. évi emelt szintű vizsga 994, míg a 2006. évi emelt szintű vizsga 810 vizsgázó adatát tartalmazza.

[4] A középszintű érettségi feladatsorok, javítási útmutatók, hanganyagok az Oktatási és Kulturális Minisztérium honlapján elérhetők:
http://www.okm.gov.hu/main.php?folderID=266&articleID=7293&ctag=articlelist&iid=1.

lőnek ítélt feladat vagy item esetén így nem tudjuk biztosan meghatározni, hogy a populáció sajátosságáról vagy teszthibáról van szó (*Molnár*, 2003).

A Rasch-modellel elemző szoftverek logaritmikus transzformációt hajtanak végre az item- és személyadatokon, az ordinális skálán lévő adatokat áttranszformálják intervallumskálára, így a modell alkalmazásával lehetőség nyílik egy objektív itemnehézségi sorrend felállítására, amely egybevethető a vizsgált személyek képességeinek értékeivel. „A vizsgázók egy csoportjában kipróbált itemek biztonsággal megfelelőnek vagy alkalmatlannak ítélhetők egy teljesen más csoport vonatkozásában is. Mindez olyan előny, amelyet a klasszikus elemzési módszerek nem képesek nyújtani" (*Nikolov*, *Pércsich* és *Szabó*, 2000. 10. o.).

A továbbiakban olyan kutatási kérdéseket fogalmazunk meg, amelyekre a Rasch-modell alkalmazásával választ kapunk. (1) A feladatsorok minősége szempontjából meghatározzuk a feladatok, itemek nehézségi szintjét, és azt vizsgáljuk, hogy a feladatkészítők által megállapított nehézségi sorrend mennyire esik egybe az empirikus nehézség növekedésével. (2) A feladatok nehézségével kapcsolatosan egyrészt azt elemezzük, hol van ugrás az egyes itemek nehézségi indexei között, másrészt azt, hogy a feladatok megfelelő nehézségűek-e a vizsgázók számára, vagyis az itemnehézségi indexek mennyire fedik le a vizsgázók nyelvi képességszintjei által meghatározott képességskála-intervallumot. (3) A feladatfejlesztés szempontjából arra keresünk választ, hogy melyik itemet milyen okból lehet elhagyni a vizsgált füzetekből, mert nem volt hasznos és törölhető. A vizsgálat célja a feladatsorok, a feladatok és az itemek működésének teljesebb megismerése, a nehézségi szintek meghatározása, az abból levonható következtetések megfogalmazása.

Az elemzés kérdéseinek megválaszolására és a célok elérésére az adatfájlban úgy rögzítettük az adatokat, hogy a helyes válasz 1, a helytelen 0 pontot ért. A nyelvhelyesség füzetben a feleletválasztós feladatokban a disztraktorokat is kódoltuk, így ezek elemzésére szintén lehetőség nyílt. A tanulmányban bemutatott elemzések, ábrák a ConQuest (*Wu*, *Adams* és *Wilson*, 1998) szoftverrel készültek.

A Rasch-modellel kapott adatok értelmezése előtt megvizsgáltuk, hogy az egyes feladatsorok mennyire megbízhatóak. Az 1. táblázatban az adatokat a nyerspontszám alapján számoltuk ki, amelynek maximális értékét minden feladatsor esetén az itemszám adja.

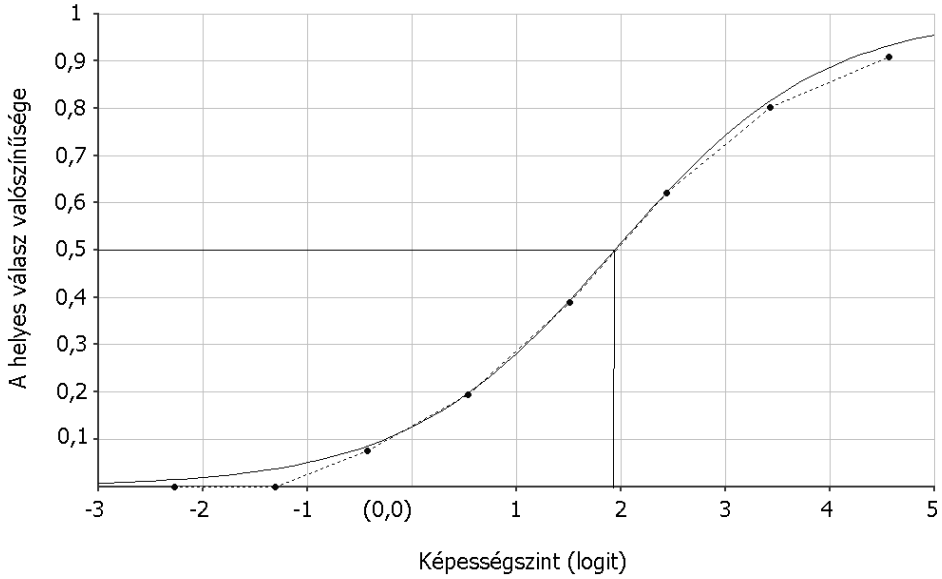
1. táblázat. A feladatlapok átlaga, szórása és Cronbach-α-ja a nyerspontok alapján (N=1092)

	Olvasott szöveg értése (itemszám=25)	*Hallott szöveg értése (itemszám=24)*	*Nyelvhelyesség (itemszám=30)*
Átlag	17,88	16,50	19,57
Szórás	4,31	4,17	5,17
Cronbach-α	0,80	0,81	0,81

Az itemek nehézségi indexe

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. o. idézi *Molnár*, 2006. 106. o.). Ennek az elgondolásnak megfelelően a Rasch-modell segítségével olyan itemjelleggörbéket rajzolhatunk, amelyek egy itemre adott helyes válaszok valószínűségét különböző képességszintek mellett írják le és megadják, hogy az egyes képességszintű diákok milyen valószínűséggel válaszolják meg helyesen az adott itemet.

Az 1. ábra példaként egy jól működő item karakterisztikus görbéjét mutatja. A vízszintes tengelyen a képességszinteket, a függőleges tengelyen az ezekhez tartozó megoldási valószínűségeket (0 és 1 közötti érték) ábrázoltuk. Az 1. ábráról leolvashatjuk, hogy mely képességszintű vizsgázók oldják meg nagyobb valószínűséggel az itemet, valamint az item mennyire különíti el egymástól a különböző képességűeket. A helyes válasznak még igen alacsony képességszint értékeknél is van valószínűsége, és a tévedésnek is van valószínűsége még az igen magas képességszint értékek mellett is (*Verhelst*, 2004). Az itemek nehézségét az a képességszint reprezentálja, ahol a helyes megoldás valószínűsége 50 százalék. Az 1. ábrán jelöltük ezt az értéket is (1,94). A képességszintek és az itemnehézségi paraméterek közös skálájának egységét logitnak nevezzük.



1. ábra
A hallott szöveg értése feladatsor 15. itemének jelleggörbéje

A továbbiakban az egyes feladatok itemeinek nehézségi indexét jellemezzük, amihez a 2. táblázatban zárójelben megadjuk a jellemzett feladatok típusait is. A feladatlapokat egy adatbázisban, egy skálára hozva elemeztük, így közvetlenül összehasonlítható a feladatok itemeinek nehézsége, skálaterjedelme és szórása.

2. táblázat. A feladatlapok nehézségi indexének főbb jellemzői

Feladatlap	*Feladat*	*Itemek száma*	*Skála logitok*		*Középérték logitok*	*Skála terjedelme*	*Szórás*
			Alsó	*Felső*			
Olvasott szöveg értése	1. (táblázat kitöltése)	8	-2,78	0,41	-1,17	3,19	0,96
	2. (interjúkérdések párosítása válasszal)	7	-0,14	2,91	1,08	3,05	1,11
	3. (rövid választ igénylő nyitott kérdések)	10	-1,60	1,41	0,18	3,01	0,99
Hallott szöveg értése	1. (igaz/hamis állítás)	7	-2,54	0,79	-0,97	3,33	1,03
	2. (rövid választ igénylő nyitott kérdések)	7	-2,48	2,52	0,03	5,00	2,04
	3. (hiányos mondatok kiegészítése)	10	-1,59	1,95	0,66	3,54	1,36
Nyelv-helyesség	1. (mondatalapú feleletválasztás)	7	-2,20	1,18	-0,60	3,38	1,06
	2. (szövegalapú feleletválasztás)	8	-1,49	0,93	-0,34	2,42	0,81
	3. (igék ragozott alakjainak beillesztése)	9	-1,11	1,07	0,13	2,18	0,82
	4. (szöveg kiegészítése elöljárószavakkal)	6	0,04	2,34	1,22	2,38	0,77

A 2. táblázat bemutatja, hogy a feladatkészítők által megállapított nehézségi sorrend mennyire esik egybe a feladatok empirikus nehézségének növekedésével[5]. Ez többnyire akkor teljesül, ha a feladatok itemeinek alsó és középérték logitértékei egyre növekednek. Ennek a kritériumnak a hallott szöveg értése és a nyelvhelyesség feladatsor felel meg leginkább, tehát a feladatkészítők megfelelően határozták meg ezekben a füzetekben a feladatok nehézségi szintjét. Ugyanakkor ezek az adatok azt is jelzik, hogy az olvasott szöveg értése feladatsor második feladata a harmadiknál nehezebb, így nem teljesül az a szándék, hogy a harmadik feladat legyen a legnehezebb. E jelenség oka a második feladat típusában keresendő. Ebben a feladatban a vizsgázóknak egy interjú válaszaihoz kellett az előre megadott kérdéseket hozzárendelni. A feladat itemnehézségi indexeinek vizsgálata alapján két kiugróan nehéz (12. és 13.) itemet azonosíthatunk, amelyek elemzése alapján (*Einhorn*, 2008) megállapítható, hogy a kérdésre adott válasz tartalmilag nem kötődik eléggé a kérdéshez, így ezeket az itemeket ki kell hagyni egy tesztfejlesztés során.

A 2. táblázat adatai a hallott szöveg értése feladatsor második feladatának belső instabilitására is utalnak, mivel az itemek nehézségi indexei nagy szóródást mutatnak, a skálaterjedelem 5 logitegységnyi. Ennek feltehetően az az oka, hogy a rövid szöveghez túl sok item készült és a két legnehezebb (8. és 12.) item megoldásakor számokat kellett a vizsgázóknak beírni, amelyeknek megértése általában nehézséget okoz. Mindkét feladatnál a zárójelben megadott itemek elhelyezkedését a 2. ábrán tanulmányozhatjuk.

[5] A tesztfüzeteket úgy állítják össze, hogy az első feladat legyen a legkönnyebb az utolsó pedig a legnehezebb.

A 2. táblázat alapján a nyelvhelyesség feladatsor első feladatában a legnehezebb item nehézségi indexe magasabb, mint a második és harmadik feladat legnehezebb iteme. Ez a 3. item, amelynek viselkedését a továbbiakban részletesen elemezzük (lásd 3. ábra).

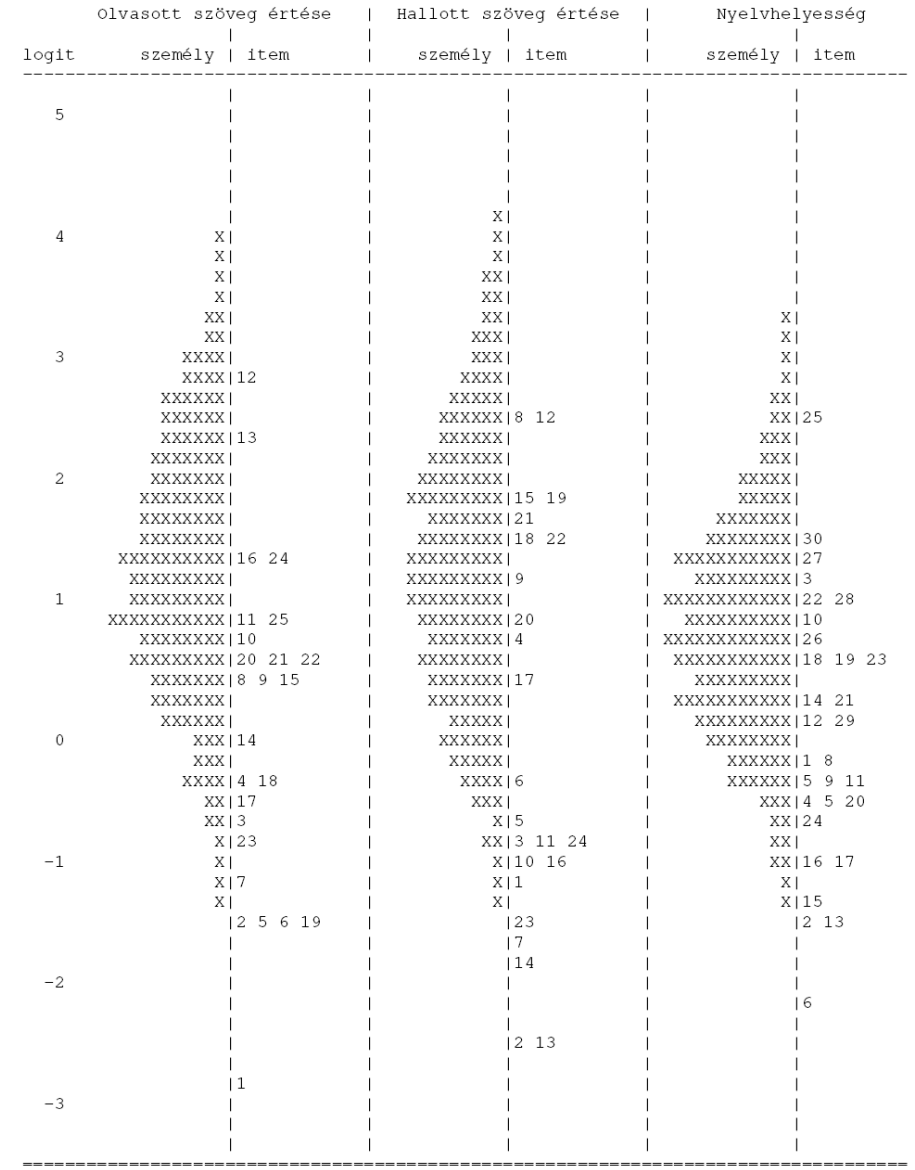
Az itemek nehézségi szintjeinek és a vizsgázók képességszintjeinek összekapcsolása

A Rasch-modell segítségével a 2. táblázatban jellemzett feladatok itemeinek nehézségét és a mintába került diákok képességszintjét a minta-, illetve itemtérkép (*map of persons ability/ item's difficulty map*) segítségével közös képességskálán ábrázolhatjuk, így össze tudjuk hasonlítani a diákok képességszintjét a feladatlapon szereplő itemek 50 százalékos valószínűséggel történő megoldásához szükséges képességszintekkel. Mivel a Rasch-modellel elemző szoftverek az adatokat intervallumskálára transzformálják, a térképről leolvasható, hogy az adott item mennyivel könnyebb-nehezebb, illetve a vizsgált személy mennyivel jobb-rosszabb képességű. A térkép az itemnehézségi indexeket és a képességparamétereket közös logitskálán ábrázolja, amely nem határozza meg a képességszintek és a nehézségi indexek abszolút helyét, hanem felállítja a relatív távolságokat a képességszinteken és a nehézségi indexeken belül, továbbá a képességszintek és a nehézségi indexek között.

A minta- és itemtérképek többféleképpen kirajzolhatók (lásd pl. *Molnár*, 2003, 2005, 2006; *Molnár* és *Józsa*, 2005), esetünkben feladatlaponként vagy együtt, több dimenzióban ábrázolhatjuk ezeket. Mindhárom feladatlapot ugyanazon vizsgázók töltötték ki és célunk az eredmények összehasonlítása, így a 2. ábrán az olvasott, a hallott szöveg értése, valamint a nyelvhelyesség feladatsor minta- és itemtérképét közös skálára kalibrálva, három dimenzióban mutatjuk.

A 2. ábrán minden egyes 'x' hét tanulót reprezentál. Fontos utalni arra, hogy az ábrázolt 'x'-eken kívül is találhatók vizsgázók a képességskála magasabb, illetve alacsonyabb tartományában, csak hétnél kevesebben vannak, így az ábra nem tartalmazza az adatokat. Az elemzésben alkalmazott ConQuest program (*Wu*, *Adams* és *Wilson*, 1998) a képességszintek átlagát nullának veszi, ezért a negatív számok átlag alatti képességet jelölnek. A 2. ábra alapján a vizsgált személyek képessége mindhárom füzet esetén rendkívül tág határok között mozog, jelentős különbségek vannak az egyes vizsgázók között. Ez a megállapítás nem meglepő, hiszen a mintában nagyon eltérő nyelvtudásszintű vizsgázók szerepelnek. A három feladatlap képességeloszlását összehasonlítva a hallott és az olvasott szöveg értése feladatsor képességparamétereinek eloszlása nagyon hasonló képet mutat.

A közös logitskálán ábrázolt feladatlapok minta- és itemtérképe alapján megállapítható, hogy az adott feladatlap itemeinek nehézsége mennyire felel meg a vizsgázók nyelvi szintjének, az itemnehézségi indexek mennyire jól fedik le a vizsgázók nyelvi szintjei által meghatározott képességskála-intervallumot. A mintához jól illesztett teszt esetén a személyek képességparamétereit jelző 'x'-ek és az itemeket reprezentáló számok egymással párhuzamosan futnak. Ennek a kritériumnak a három feladatlap közül a nyelvhelyesség füzet felel meg a leginkább, hiszen a feladatok itemei jól kiegészítik egymást és így többnyire megfelelően fedik le a vizsgázók képességszintjét.



2. ábra

Az olvasott és hallott szöveg értése, a nyelvhelyesség feladatlap minta- és itemtérképe[6]

[6] A ConQuest program (*Wu, Adams* és *Wilson*, 1998) a többdimenziós ábrázolás során egymás mellé helyezi először a képességparamétereket majd az itemeket. Ezt az ábrát szerkesztettük, mivel füzetenként elkülönítve vizuálisan jobban összehasonlítható a minta- és itemtérkép.

Ugyanakkor az olvasott szöveg értése feladatlapon három helyen találunk olyan itemcsoportosulást, amelyek azonos képességszintet mérnek, így az itemek nem fedik le a teljes képességskála-intervallumot. A két kiugróan nehéz (12. és 13.) itemet a 2. táblázat során már elemeztük. A hallott szöveg értése feladatsorra egyrészt jellemző, hogy az átlagos képességszintű diákok képességszintjét (logitérték=0) mérő itemek hiányoznak a tesztből, másrészt két egymástól jól elkülöníthető itemcsoportot találunk, amelyek egyik része az átlagos képességszint alatt, a másik az e fölötti képességszintű vizsgázókat méri. A három feladatlap közös jellemzője továbbá, hogy van egy magas képességszint, amit a feladatsorok már nem mérnek.

Az itemek megbízhatósági mutatói

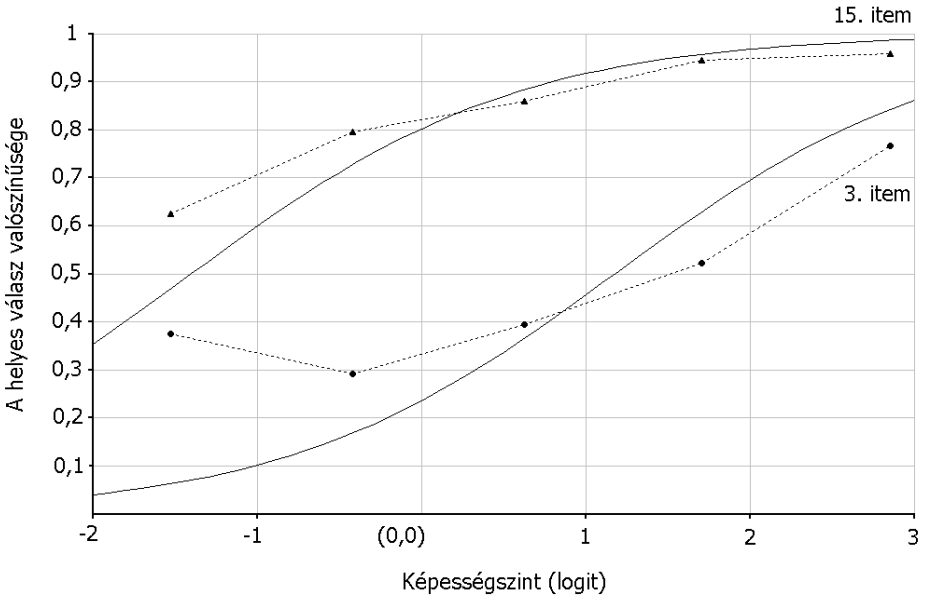
Az itemek megbízhatósági mutatói megmutatják, hogyan lehetne a feladatlapokat továbbfejleszteni, illetve azt is, hogy melyik itemet lehet elhagyni vagy módosítani. A Rasch-modellel történő tesztfejlesztés során meg kell vizsgálni a feladatok itemeinek diszkriminációs indexét, modellilleszkedését, és a nem megfelelően működő itemek karakterisztikus görbéjét.

A 2. ábra alapján az olvasott és a hallott szöveg értése füzetben öt, míg a nyelvhelyesség feladatsorban három olyan item található, amelyik logitértéke alacsonyabb, mint az összes diák képességparamétere. Ez arra utal, hogy annak a valószínűsége, hogy ezeket az itemeket az adott mintában valaki megoldja, nagyobb, mint 50 százalék, továbbá magas annak a valószínűsége, hogy az átlagos képességűek (logitérték=0) megoldják ezeket az itemeket. Ha ezeknek az itemeknek a diszkriminációs indexét megvizsgáljuk, akkor 0,1 és 0,2 közötti értékeket találunk, amelyek azt mutatják, hogy ezek az itemek nem jól különítik el egymástól a mintába bekerült jobb és gyengébb képességű vizsgázókat. Alacsony diszkriminációs indexű itemet az átlagos képességszint fölött is találunk. Példaként a 3. ábrán közösen ábrázoltuk a nyelvhelyesség füzet 3. és 15. itemét, amelyek csak nehézségi indexükben különböznek, a diszkriminációs indexük egyaránt alacsony (0,2). Mindkét item esetén látható, hogy nem rajzolódik ki a teljes logisztikus görbe, mint az 1. ábrán. Az alacsony diszkriminációs indexű nehéz item (3.) azt jelzi, hogy az itemet kevesen, de az egész teszten az átlagnál jobb és gyengébb teljesítményt nyújtó vizsgázók közel azonos valószínűséggel oldják meg.

A 3. item az alacsony diszkriminációs index mellett egy problémára is utal. Jól látható, hogy az empirikus item karakterisztikus görbe (szaggatott vonal) laposabb, mint az elméleti görbe (folytonos vonal), tehát az adatok nem illeszkednek az elvárt, előre jelzett modellhez.

Az itemek működése és a feladatok esetleges továbbfejlesztése során fontos megvizsgálni, hogy az adatok mennyire követik a modellt. Az illeszkedésvizsgálat arról ad információt, hogy mennyire illeszkedik az adott item a többi közé, így rávilágít a problémás itemekre. A három feladatlapon nyolc nem illeszkedő itemet találtunk, amelyek paramétereit a 3. táblázat mutatja. Az infit (súlyozott) paraméter „azokra a válaszokra vonatkozik, melyek a vizsgázó képességéhez (a feladat nehézségéhez) közel vannak, és ezért jellemzőnek, tipikusnak mondhatók” (*Dávid*, 2005. 13. o.). Az adatok illeszkedése esetén az infit paraméter értéke 1-hez tart, ekkor az itemek diszkrimináló ereje közel

azonos (*Molnár*, 2006). Mivel az elfogadható értékek sávja függ a minta elemszámától, így nem határozható meg előre az elfogadási sáv, hanem a vizsgálat céljától és alkalmazásától függően kell mérlegelni (*Wu*, 2006b idézi *Molnár*, 2006. 112. o.; *Linacre*, 2002b idézi *Eckes*, 2003. 54. o.). Irányelvként megállapítható, hogy például egy 2000 fős minta esetén az elfogadható értékek 0,94 és 1,06 között ingadozhatnak (*Molnár*, 2006). Az 1,06 fölötti értékű itemek nem illeszkednek a modellbe, tehát azt jelzik, hogy mást mérnek.



3. ábra
A nyelvhelyesség füzet 3. és 15. itemének jelleggörbéje

A 2. ábra és a 3. táblázat alapján jól látszanak a tesztfejlesztés lehetőségei. Például az olvasott szöveg értése feladatlap minta- és itemtérképe (2. ábra) mutatja, hogy a harmadik feladathoz tartozó 20., 21. és 22. item azonos képességszintet mér. Mivel célunk az, hogy a képességparamétereket minél jobban lefedje a teszt, indokolt a nem illeszkedő 21. item cseréje.

A nem illeszkedő itemek viselkedését *Nikolov*, *Pércsich* és *Szabó* (2000) alapján a feladat önálló, elkülönített elemzésével vizsgálhatjuk, amelyhez szükséges a feladatok ismerete is. Ez már túlmutat a tanulmány célkitűzésein, ezért csak néhány szempontra térünk ki. A hallott szöveg értése feladatsor első feladatának külön elemzése semmilyen rendellenességet nem mutatott, nem találtunk nem illeszkedő itemeket. A feladat egyedül jól funkcionál, de a füzetben lévő két feladattal együtt már kevésbé. Ennek feltehetően az oka, hogy a feladat inkább globális szövegértést mér, a vizsgázónak a feladattípusból adódóan (igaz-hamis állítás eldöntése) kevesebb tevékenysége van, magas a találgatás

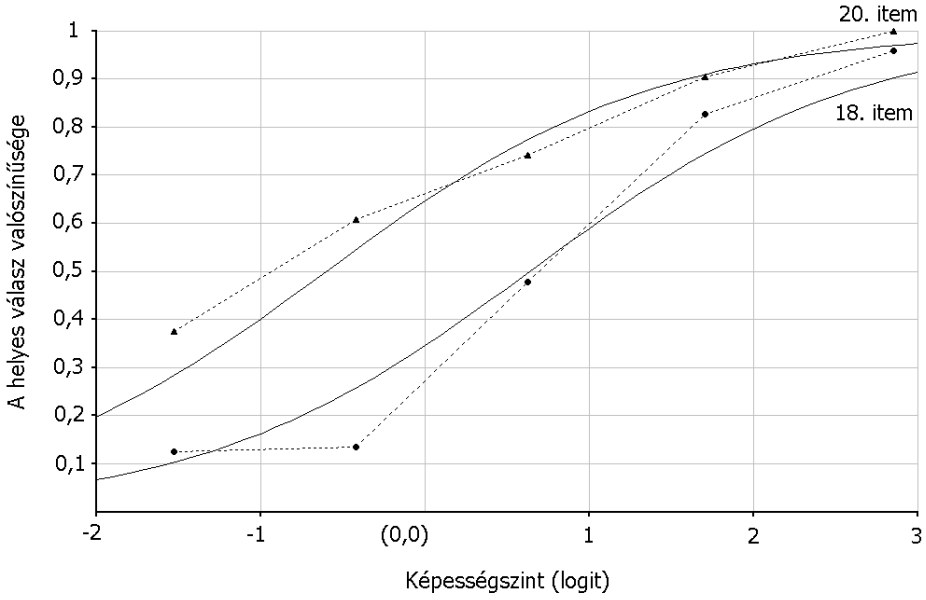
lehetősége, míg a másik kettő feladatban egyszerre kell figyelnie a hallott szövegre, valamint a kérdésekre (2. feladat), vagy a kiegészítendő mondatokra (3. feladat) és közben még írnia is kell.

3. táblázat. A nem illeszkedő itemek paraméterének jellemzői

Feladat	*Itemszám*	*Nehézség*	*Diszkrimináció*	*Infit paraméter*
Olvasott szöveg értése 3. feladat	21	0,54	0,22	1,24
Hallott szöveg értése 1. feladat	3	-0,95	0,33	1,09
	6	-0,45	0,29	1,16
Nyelvhelyesség 1. feladat	2	-1,51	0,14	1,09
	3	1,18	0,21	1,21
Nyelvhelyesség 3. feladat	20	-0,60	0,31	1,07
	21	0,34	0,35	1,07
Nyelvhelyesség 4. feladat	26	0,81	0,28	1,14

A nyelvhelyesség feladatsor első feladatának elemzése sok problémára utal (vö. 2. ábra, 2. és 3. táblázat). Ezt a feladatot össze tudjuk hasonlítani a második feladattal, ami szintén feleletválasztós, csak nem mondat-, hanem szövegalapú. Közös a két feladatban, hogy mindkettő az alacsonyabb (A2) szinthez készült, különbség viszont, hogy a szövegalapú feleletválasztós feladat itemeinek megbízhatósági mutatói jobbak. Az első feladat problémája, hogy a kontextus hiánya alapvetően befolyásolja az itemek megoldottságát és működését. Vizsgálatunk is alátámasztja a mondatalapú feleletválasztós feladat törlésére tett javaslatot a választható feladattípusok közül (*Einhorn*, 2008).

A nyelvhelyesség füzet harmadik feladatának elkülönített elemzése során a 3. táblázatban jelölt két, nem illeszkedő item mellett négy olyan itemet is találtunk, amelyek infit paraméterei 0,9 körüli értékűek. A továbbiakban példaként ebből a feladatból két item (18. és 20.) karakterisztikus görbéjét ábrázoljuk közösen (4. ábra). A 18. item jelleggörbéje az elvártnál meredekebb, és azt mutatja, hogy egy bizonyos képességszintig alacsony annak a valószínűsége, hogy egy személy helyesen válaszolja meg az itemet, majd megnő az item megoldásának valószínűsége, és többnyire mindenki jól válaszol. Ez azt jelenti, hogy az item túlilleszkedik, mivel diszkriminációs indexe magas (0,59), ám ez az itemműködés általában nem jelent problémát (*Dávid*, 2006; *Nikolov*, *Pércsich* és *Szabó*, 2000). A 20. item jelleggörbéje a 3. itemhez (3. ábra) hasonlóan az elvártnál laposabb, tehát az item nem illeszkedik a többi közé. Ez az itemműködés azért problematikus, mert azt jelzi, hogy nagy számban vált ki váratlan válaszokat a vizsgált személyektől. Ebben a feladatban a vizsgázóknak főnévi igeneveket kellett megfelelő igealakban behelyezni a szövegbe. A vizsgázók elemzett megoldásai (*Einhorn*, 2008) tanítási-tanulási problémát feltételeznek, mert az érettségizőknek gyakran nem az okozott problémát, hogy a megfelelő igeidőt válasszák, hanem nem ismerték a helyes igealakokat.



4. ábra
A nyelvhelyesség füzet 18. és 20. itemének jelleggörbéje[7]

Az adatbázisok összekapcsolásának, a feladatbank kiépítésének problémái

A minta- és itemtérkép, valamint az itemek megbízhatóságának vizsgálata alapján számos információ szerezhető az egyes feladatok itemeinek nehézségéről, a fejlesztés lehetőségeiről, azonban több kérdés is felmerül, amelyre nem kapunk választ.

Az egyik kérdés a középszintű érettségi szintjeit érinti. A középszintű vizsga a Referenciakeret két szintjét méri, azonban a minta- és itemtérképről (2. ábra) nem tudjuk megállapítani, hogy mely vizsgázó tudása minősül A2, mely érettségiző tudása B1 szintűnek, továbbá, mely itemek nehézségi indexét lehet az egyik, illetve másik szinthez sorolni. A 2. táblázatban jellemzett itemnehézségi indexek, valamint a minta- és itemtérkép (2. ábra) mutatja, hogy a könnyebb feladatok általában az átlagos képességszint alatt helyezkednek el, de azt nem tudjuk, hol kell meghúzni a határokat, nem különül el egymástól a kétféle tudásszint.

[7] A két item diszkriminációs indexe eltérő (0,59 és 0,31). A Rasch-modell nem határozza meg a diszkrimináció abszolút nagyságát, így a modell képileg az itemeket azonos meredekségű logisztikus görbével ábrázolja. A jobban diszkrimináló itemek azonban jobban széthúzzák a képességparaméter-értékeket (*Molnár*, 2006).

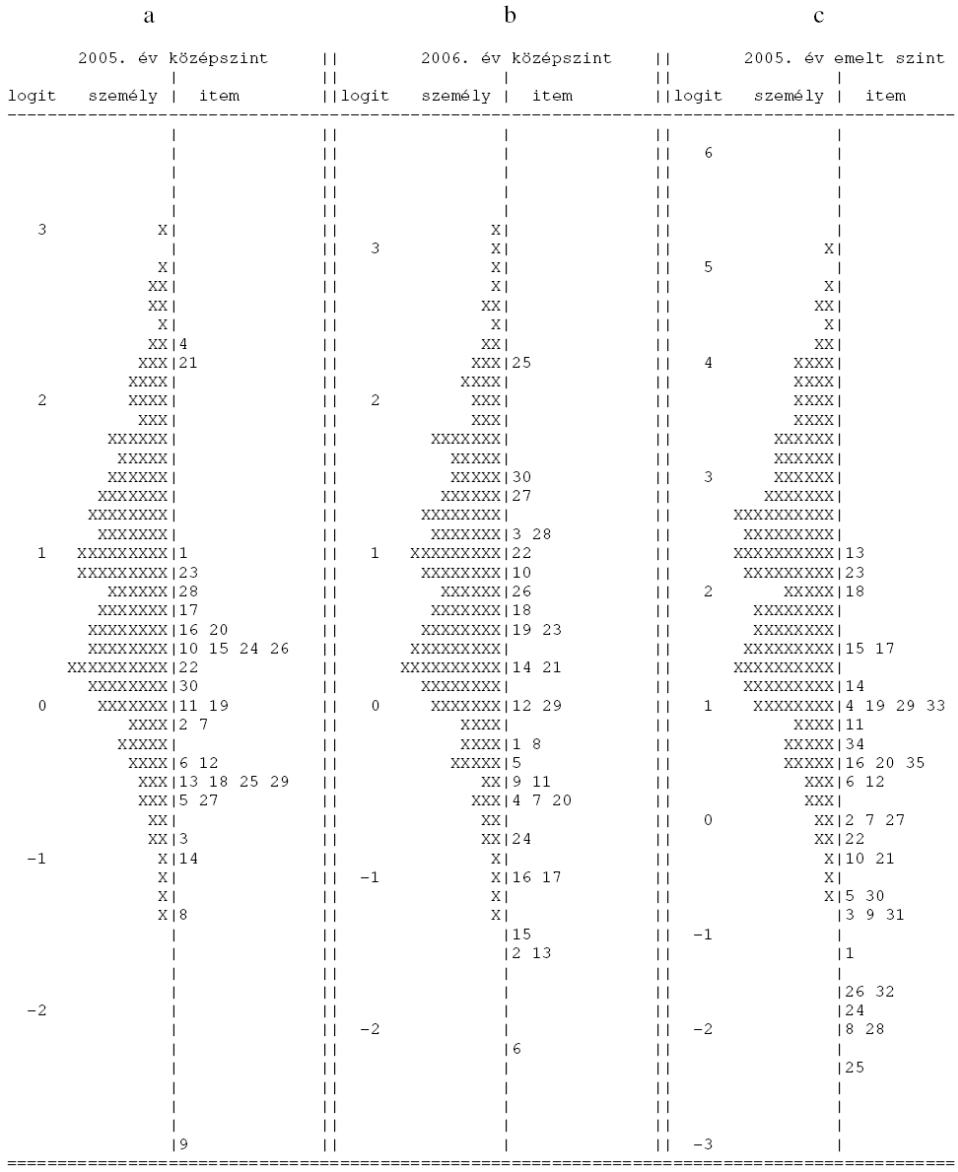
A további kérdések tárgyalásánál a nyelvhelyesség feladatsor példáján mutatjuk be a vizsgaszintek empirikus vizsgálatának szükségességét. Az 5.a, 5.b és 5.c ábrán egymás mellé vetítve mutatjuk a 2005. évi, az előzőekben részletesen elemezett 2006. évi középszintű, valamint a 2005. évi emelt szintű német érettségi nyelvhelyesség feladatlap minta- és itemtérképét. Az adatokat nem skálázhattuk együtt, mint a 2. ábrán, mivel nincs olyan személy vagy item, ami összekötné az adatbázisokat, így az összehasonlítás nem releváns. Nem mondhatjuk például, hogy a 2005. évi érettségi 2. iteme ugyanolyan nehézségű, mint a 2006. évi nyelvhelyesség feladatlap 1. iteme, mivel a két adatbázis között semmiféle kapcsolat nincs. Ezért választottuk el a három térképet szaggatott vonallal. A három ábra egymás mellé vetítése csak a problémafelvetést szemlélteti.

Az 5.a és 5.b ábra alapján megfogalmazható az a kérdés, hogy a feladatlapok nehézségi szintje mennyire állandó, mennyire tekinthetők ekvivalensnek? Ennek tisztázása főként a Referenciakeret két szintjét mérő középszint esetén fontos, mivel már minimális szinteltolódás esetén is jelentősen eltérő nehézségű feladatsor jöhet létre. Bár közvetlenül nem hasonlítható össze a két feladatlap minta- és itemtérképe, az azonban megállapítható, hogy a személyek képességparamétere hasonló eloszlású a két évben. A 2006. évi német nyelvű érettségi esetén ennél a feladatlapnál az itemek jobban fedik le a vizsgázók képességparaméterei által meghatározott képességskála-intervallumot, mivel az 5. a ábrán két helyen találunk négy olyan itemet, amelyek azonos képességszintet mérnek. A tesztfejlesztés során erre a problémára megoldás lehet az alacsony megbízhatósági mutatóval rendelkező itemek cseréje. A 2005. évi emelt szintű nyelvhelyesség feladatlap minta- és itemtérképéről (5.c ábra) megállapítható, hogy a feladatlap a vizsgázóknak túl könnyű volt. Az itemnehézségi indexek fele helyezkedik el a 0 logitegység alatt.

Az 5.a, 5.b és 5.c ábra alapján felmerül az a kérdés, hogy a középszinten magasabb teljesítményt elérők milyen valószínűséggel oldanák meg az emelt szintű feladatokat, vagyis hogyan viszonyul egymáshoz a közép- és az emelt szintű feladatsorok nehézségi szintje? Az 5.c ábra egy másik problémára is utal. A magasabb képességtartományban az emelt szintű feladatlap nem mér.

További kérdésként fogalmazható meg, hogy valóban B2 szintet mérnek-e az emelt szintű feladatok? Amennyiben a feladatok B2 szintűek, akkor indokolt lehet a vizsgakövetelmények módosítása és a feladatok nehézségi szintjének emelése C1 szintre. Amennyiben a feladatok a B2 szint alatti szintet mérnek, akkor szükséges a vizsgafeladatokat a követelményekben megadott szinthez igazítani. Ezek a hipotézisek is jelzik, hogy az emelt szintű feladatsor esetén is szükséges megvizsgálni a Referenciakerethez történő illeszkedést.

A Rasch-modell segítségével a fenti kérdésekre tudunk válaszolni, mivel a modell egyik fontos tulajdonsága – amit a PISA-vizsgálatokban (*OECD*, 2005) is alkalmaznak –, hogy „ha ismerjük egy diák képességszintjét, meg tudjuk mondani, hogy milyen valószínűséggel oldana meg olyan itemet, amelynek nehézségi indexe értelmezhető a közös képességskálán, anélkül, hogy a diáknak a valóságban meg kellene oldani azt” (*Molnár*, 2006. 103–104. o.). A Rasch-modell a feladatokat összekötő horgonyitemek (*anchor items*) segítségével lehetőséget teremt a feladatsorok együttes elemzésére, a feladatok egy skálára hozására, tehát közvetlenül összehasonlíthatóvá válik az item nehézsége és a diákok képességszintje (*Molnár*, 2003).



5. a, 5. b és 5. c ábra
A 2005. és a 2006. évi középszintű és a 2005. évi emelt szintű német érettségi nyelvhelyesség feladatlap minta- és itemtérképe[8]

[8] Az 5.a és 5.b ábrán minden 'x' hét személyt jelöl, míg az 5.c ábrán hat vizsgázót reprezentál.

A feladatok Referenciakerethez történő illeszkedésének vizsgálatához referenciaitemeket tartalmazó feladatokra van szükség. „Referencia olyan item lehet, amelynek nehézségi értékei vagy a referenciaskálához kapcsolt, „kihorgonyzott" mérésben keletkeztek, vagy attól nem függő mérés esetében a bemérendő és a referenciaskála skálatulajdonságainak kell egyeznie" (*Dávid*, 2005. 290. o.). A referenciaitemek segítségével meghatározhatjuk azt a nehézségi szintet, amelyhez a program a többi item nehézségének meghatározásakor viszonyít és így megállapíthatjuk, hogy a vizsgafeladatok itemei mennyivel könnyebbek-nehezebbek a referenciaitemeknél. Ebben az eljárásban is szükséges, hogy legyenek közös itemek (*Molnár*, 2006).

Az európai projektekben a Referenciakeret deskriptorainak, nyelvi szintjeinek bemérése során a DIALANG[9] értékelő rendszer kialakításában és az EBAFLS-projektben alkalmazták ezeket az eljárásokat. A DIALANG egy olyan online adaptív nyelvi értékelési rendszer, amely a Referenciakeret diagnosztikai célokra történő alkalmazásaként jött létre. A rendszer a Referenciakeret deskriptorai alapján kalibrált önértékelési skálákat és nyelvi teszteket tartalmaz, és ezek alapján nyújt diagnosztikus információkat. A rendszer eredményként megadja, hogy a tanulók az olvasás és a hallás utáni szövegértés, az írás, a nyelvtan és szókincs területén milyen szinten állnak, valamint visszajelzést ad készségszintjük erős és gyenge pontjairól, továbbá tanácsot ad a készségek fejlesztéséhez (*KER C függelék*, 2002; *Alderson* és *Huhta*, 2005). Az EBAFLS-projekt alapvető célja a Referenciakeret alapján referenciafeladat-bank létrehozása. A projektben nyolc európai ország, köztük Magyarország részvételével francia, német és angol nyelvhez, olvasott és hallott szöveg értése készségekhez B1 szintre hoznak létre referenciafeladatokat. A létrejövő feladatbankok célja az itemek pszichometriai jellemzőinek és a Referenciakeret szintjeinek megadásával elősegíteni, hogy a különböző nyelvtudást mérő vizsgákat a Referenciakeret szintjeihez empirikusan tudják validálni (*CITO*, 2008).

A validálási folyamatban és a vizsgafeladatok nehézségi szintjének vizsgálatában a horgonyitemek és a Referenciakeret szintjeihez kapcsolt, kalibrált referenciaitemek alkalmazásával számos módszertani kérdés is felmerül. Az egyik problémakör a referenciafeladatok megfelelő kiválasztása. Az EBAFLS-projekt kutatási beszámolója (*CITO*, 2008) egyértelműen utal a feladatbank alkalmazásának korlátaira is, mivel sok item különböző működést (*differential item functioning*) mutatott az egyes országokban. Ez az itemműködés azt jelenti, hogy a különböző országokból származó tanulók más valószínűséggel oldották meg az itemeket. A helyes válasz valószínűsége nemcsak a mérendő tulajdonságtól, hanem más tényezőktől is függ. Ez az itemműködés jelezheti a tanterv és a nyelvoktatatási gyakorlat közötti különbségeket, az eltérő mérési és értékelési hagyományokat (egy- vagy kétnyelvű vizsgák alkalmazása). Befolyásoló tényező még a feladattípusok előzetes ismertsége, továbbá az is, hogy az adott nyelv struktúrája menynyire van közel a mérendő idegen nyelvhez. A beszámoló egyértelműen jelzi, hogy a különböző itemműködés okainak vizsgálatához további kutatásokra van szükség a nyelvi mérés és értékelés területén. A problémákat úgy küszöbölik ki, hogy az egyes országok szintjén is megadják az itemek nehézségi indexét, diszkriminációs indexét. A másik fontos kérdéskör, hogy mennyi horgony- és referenciaitemre van szükség, amelyekkel az

[9] A program 14 európai nyelven a http://www.dialang.org honlapról tölthető le.

adatbázisokat megfelelően össze lehet kötni (*Szabó*, 2006), illetve megteremtik a kapcsolatot a Referenciakeret szintrendszerével.

A továbbiakban bemutatjuk, hogy ezeknek a korlátoknak az ismeretében hogyan lehet létrehozni a német nyelvi érettségi feladatsoraiból a vizsgált adatbázisok alapján feladatbankot. Az EBAFLS-projekt ajánlásainak megfelelően a legfontosabb lépések a következők:

I. Az érettségi vizsgafeladatok ekvivalenciájának, nehézségi szintjének vizsgálata a Referenciakeret alapján:
   1. Az adatbázisban lévő feladatok nehézségi szintjének vizsgálata.
   2. Érettségi és referenciafeladatok kiválasztása.
   3. Tesztfüzetek kialakítása.
   4. Tesztfüzetek bemérése.
   5. Adatok bevitele, adatbázis létrehozása.
   6. IRT-alapú itemelemzés és –szintezés.

II. Az érettségi feladatokból feladatbank létrehozása:
   7. Feladatok kiválasztása.
   8. A feladatok jellemzőinek leírása (a Dutch grid-projekt alapján[10] *Alderson* és mtsai, 2004).
   9. Ajánlások megfogalmazása az érettségi vizsgafeladatok fejlesztéséhez.

Az első fázisban, az idegen nyelvi érettségi vizsgaszintjeinek vizsgálatában a különböző vizsgaidőszakokban keletkező feladatokból horgonyitemek segítésével és újabb felmérésben történő bemérésével meg tudjuk teremteni a különböző vizsgafeladatok itemszintű adatbázisai közötti kapcsolatot, és a Rasch-modell alkalmazásával egy adatbázisban tudjuk elemezni az adatokat. Ennek a kutatásnak a célja az azonos nyelvi konstruktumot mérő, különböző vizsgaidőszakban keletkezett feladatsorok nehézségének, állandóságának és ekvivalenciájának vizsgálata. A Referenciakeret szintjeihez való illeszkedést az európai projektekből származó referenciafeladatokkal ellenőrizhetnénk. Egy ilyen kutatás lebonyolítása és az eredmények értelmezése során meg tudnánk állapítani, hogy az idegen nyelvi érettségi vizsgafeladatainak két szintje mennyire igazodik a vizsgázók nyelvtudásának reális szintjéhez. A referenciafeladatok alkalmazásával választ kaphatunk arra a kérdésre is, hogy a vizsgafeladatok szintje hogyan illeszkedik a Referenciakeretben megadott nyelvtudási szintekhez. A kutatás eredménye hozzájárulhat annak ellenőrzéséhez, hogy valóban megfelelőek-e a szintekről alkotott elképzeléseink.

A második fázisban az érettségi vizsgafeladatokból feladatbank építhető ki. A feladatbankba bekerülő érettségi feladatoknak a Referenciakeret deskriptorai alapján tartalmaznia kell az itemek pszichometriai jellemzőit, tartalmát és formáját. Ebben a fázisban felhasználhatók a Dutch grid-projekt eredményei (*Alderson* és mtsai, 2004), amelyekkel a feladatokhoz tartozó szövegek és itemek jellemezhetők és leírhatók a Referenciakeret alapján. A létrejövő feladatbank viszonyítási pontként funkcionálhat a további vizsgafeladatok fejlesztésekor.

[10] A Dutch Grid a http://www.lancs.ac.uk/fss/projects/grid/ honlapon érhető el.

Összegzés

A tanulmányban egy konkrét elemzés alapján vizsgáltuk egy IRT-alapú nyelvi feladatbank létrehozásának lehetőségét, problémáit és módszertani alapkérdéseit. Empirikus vizsgálatunkban a valószínűségi modellek csoportjába tartozó Rasch-modell segítségével elemeztük a 2006. évi középszintű német nyelvi érettségi receptív készségeket és a nyelvhelyességet mérő feladatsorait.

A Rasch-modell alkalmazásával felállítottunk a vizsgált füzeteken belül egy minta- és tesztfüggetlen, objektív itemnehézségi sorrendet, amelyet egybevetettünk a mintába bekerült vizsgázók képességszintjeinek értékével. Az elemzés során azt tapasztaltuk, hogy a feladatkészítők által megállapított nehézségi sorrend többnyire megegyezett az empirikus nehézség növekedésével. A nyelvhelyesség feladatsorban a feladatok kiegészítik egymást és így többnyire megfelelően fedik le a vizsgázók képességszintjét, az olvasott és a hallott szöveg értése feladatlapokra azonban jellemző, hogy az átlagos képességszintű diákok képességszintjét (logitegység=0) mérő itemek hiányoznak. Több feladatnál két, jól elkülöníthető itemcsoportot találtunk, amelyek egyik része az átlagos képességszint alatt és felett lévő vizsgázók nyelvi képességeit mérik. Az elemzés alapján azonosítottunk egy magasabb képességszintet is, amit a feladatsorok nem mérnek. Az itemek megbízhatóságának vizsgálatakor néhány nem illeszkedő, valamint alacsony diszkriminációs indexszel rendelkező itemet találtunk, amelyek főleg az átlagos képességszint alatt helyezkedtek el.

A minta- és itemtérképek elemzése alapján további kutatási kérdéseket fogalmaztunk meg, és bemutattuk egy kutatás koncepcióját, amelyben a Rasch-modellel a receptív készségeket és a nyelvhelyességet mérő vizsgarészek feladatait lehetne vizsgálni és ezek alapján feladatbankot kiépíteni. A kutatás hozzájárulhat annak, a szakirodalomban (például *Einhorn*, 2007b; *Nikolov*, 2007) gyakran megfogalmazott kérdésnek a tisztázásához, hogy az idegen nyelvi érettségi vizsga két szintje hogyan viszonyul egymáshoz és a Referenciakeret szintrendszeréhez. A kutatás elősegítheti azt a folyamatot, amelynek eredményeként egységes skálán tudjuk elhelyezni a diákok nyelvi teljesítményét, és így az idegen nyelvi érettségi olyan évenkénti, rendszerszintű felméréssé válhat, amely egyrészt visszajelzést nyújt a közoktatás résztvevői számára, másrészt lehetővé teszi a nyelvtudásszint változásának követését az évek során.

Köszönetnyilvánítás

Köszönöm *Einhorn Ágnes*nek, hogy lehetővé tette az adatbázisok másodelemzését, továbbá *Molnár Gyöngyvér*nek és *Vidákovich Tibor*nak a szakmai támogatását az elemzés elkészítésében.

Irodalom

Alderson, J. C. (2000): Teljesítményszintek az angol nyelvi érettségi kipróbálásán. *Magyar Pedagógia*, **100**. 4. sz. 423–458.

Alderson, J. C., Clapham, C., és Wall, D. (1995): *Language test construction and evaluation*. Cambridge University Press, Cambridge.

Alderson, J. C. és mtsai (2004): *The development of specifications for item development and classification within the Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Reading and listening. Final report of the Dutch CEF Construct Project*. (http://eprints.lancs.ac.uk/44/1/final_report.pdf)

Alderson, J. C. és Huhta, A. (2005): The development of a suite of computer-based diagnostic tests based on the Common European Framework. *Language Testing*, **22**. 3. sz. 301–320.

Alderson, J. C., Nagy, E. és Öveges, E. (2000, szerk.): *English language education in Hungary. Part II. Examining hungarian learners' achievements in English*. The British Council Hungary, Budapest.

Bachman, L. F. és Palmer, A. (1996): *Language testing in practice*. Oxford University Press, Oxford.

Banerjee, J., Kaftandjieva, F., Takala, S. és Verhelst, N. (2004): *Szintillesztési módszertani segédlet. A nyelvvizsgák illesztése a Közös európai referenciakerethez című kézikönyv előzetes, kísérleti verziójához.* Nyelvvizsgát Akkreditáló Testület – PH Nyelvvizsgáztatási Akkreditációs Központ, Budapest.

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

Bárdos Jenő (2003): A nyelvtudás megítélésének korlátai. *Iskolakultúra*, **13**. 8. sz. 28–39.

Bárdos Jenő (2006): A nyelvtudás-fogalom metamorfózisai: kritikai elemzés. In: Sillye Magdolna, Kurtán Zsuzsa, Sturcz Zoltán és Wiwczaroski, T. B. (szerk.): *PORTA LINGUA – 2006. Utak és perspektívák a hazai szaknyelvoktatásban és –kutatásban*. Debreceni Egyetem, Agrártudományi Centrum, Center Print Nyomda, Debrecen. 15–22.

CITO (2008): *Building a european bank of anchor items for foreign language skills. EBAFLS. General report.* Cito B.V., Arnheim.

Dávid Gergely (2005): Nyelvvizsgaszintek validálása: lehetőségek és korlátok. In: Sillye Magdolna, Kurtán Zsuzsa, Sturcz Zoltán és Wiwczaroski, T. B. (szerk.): *PORTA LINGUA – 2005. Szakmai nyelvtudás – szaknyelvi kommunikáció cikkek, tanulmányok a hazai szaknyelvoktatásról és –kutatásról*. Debreceni Egyetem, Agrártudományi Centrum, Center Print Nyomda, Debrecen. 279–295.

Dávid Gergely (2006): *Az emelt szintű idegen nyelvi érettségi és az államilag elismert nyelvvizsgák a vizsgázói teljesítmények tükrében. Összegző tanulmány*. Kézirat. Nyelvvizsgát Akkreditáló Testület – PH Nyelvvizsgáztatási Akkreditációs Központ, Budapest.

Eckes, T. (2003): Qualitätssicherung beim TestDaF: Konzepte, Methoden, Ergebnisse. *Fremdsprachen und Hochschule*, 69. sz. 43–68.

Eckes, T. (2004): Facetten des Sprachtestens: Strenge und Konsistenz in der Beurteilung sprachlicher Leistungen. In: Wolff, A., Ostermann, T. és Chlosta, C. (szerk.): *Integration durch Sprache. Materialien Deutsch als Fremdsprache, 73.* FaDaF, Regensburg. 485–518.

Einhorn Ágnes (2004, szerk.): *Az érettségiről tanároknak 2005. Német nyelv.* Országos Közoktatási Intézet, Budapest. (http://www.okm.gov.hu/letolt/kozokt/erettsegi2005/tanaroknak/nemet/nemet_nyelv.htm)

Einhorn Ágnes (2006): A vizsgafeladat fejlesztésének folyamata és kritériumai. *Új Pedagógai Szemle*, 1. sz. 67–74.

Einhorn Ágnes (2007a): *A 2005. évi érettségi vizsga eredményeinek elemzése. Német nyelv.* Országos Közoktatási Intézet, Budapest. (http://www.oki.hu/oldal.php?tipus=cikk&kod=2005tapasztalatok-Nemet)

Einhorn Ágnes (2007b): Az idegen nyelvi érettségi vizsga reformja. In: Vágó Irén (szerk.): *Fókuszban a nyelvtanulás*. Oktatáskutató és Fejlesztő Intézet, Budapest. 73–105.

Einhorn Ágnes (2008): *A 2006. évi érettségi vizsga eredményeinek elemzése. Német nyelv.* Oktatáskutató és Fejlesztő Intézet, Budapest. (http://www.oki.hu/oldal.php?tipus=cikk&kod=2006tapasztalatok-Nemet)

Grotjahn, R. (2000): Testtheorie: Grundzüge und Anwendungen in der Praxis. In: Wolff, A. és Tanzer, H. (szerk.): *Sprache – Kultur – Politik. Beiträge der 27. Jahrestagung DaF 1999*. *Materialien Deutsch als Fremdsprache*. 53. FaDaF, Regensburg. 304–341.

KER (2002): *Közös Európai Referenciakeret: nyelvtanulás, nyelvtanítás, értékelés.* OM – PTMIK, Budapest – Pilisborosjenő.

Molnár Gyöngyvér (2003): Az ismeretek alkalmazásának vizsgálata modern tesztelméleti eszközökkel. *Magyar Pedagógia*, **103**. 4. sz. 423–446.

Molnár Gyöngyvér (2005): Az objektív mérés megvalósításának lehetősége: a Rasch-modell. *Iskolakultúra*, **15**. 3. sz. 71–80.

Molnár Gyöngyvér (2006): A Rasch-modell alkalmazása a társadalomtudományi kutatásokban. *Iskolakultúra*, **16**. 12. sz. 99–113.

Molnár Gyöngyvér (2008): A Rasch-modell kiterjesztése nem dichotóm adatok elemzése: a rangskálás és parciális kredit modell. *Iskolakultúra*, **18**. 1–2. sz. 66–77.

Molnár Gyöngyvér és Józsa Krisztinán (2006): Az olvasási képesség értékelésének tesztelméleti megközelítései. In: Józsa Krisztián (szerk.): *Az olvasási képesség fejlődése és fejlesztése.* Dinasztia Tankönyvkiadó, Budapest. 155–174.

NAT (1995): *Nemzeti alaptanterv*. Korona Kiadó, Budapest.

NAT (2003): *Nemzeti alaptanterv*. Oktatási Minisztérium, Budapest.

Nikolov Marianne (2007): A magyarországi nyelvoktatás-fejlesztési politika és annak gyakorlati megvalósulása a nemzetközi trendek tükrében. In: Vágó Irén (szerk.): *Fókuszban a nyelvtanulás*. Oktatáskutató és Fejlesztő Intézet, Budapest. 43–72.

Nikolov Marianne, Pércsich Richárd és Szabó Gábor (2000): A puding próbája. Alapszintű angol feladatok bemérésének tapasztalatai. *Modern Nyelvoktatás*, **6**. 4. sz. 3–28.

North, B., Avermaet, P.V., Figueras, N., Takala, S. és Verhelst, N. (2003): *Nyelvvizsgák szintillesztése a Közös európai referenciakerethez. Kézikönyv. Előkészítő, kísérleti változat*. Nyelvvizsgát Akkreditáló Testület – PH Nyelvvizsgáztatási Akkreditációs Központ, Budapest.

OECD (2005): The Rasch Model. In: OECD: *PISA 2003 Data Analysis Manual*. OECD, Paris. 53–70.

Petneki Katalin (2007): *Az idegen nyelvek oktatása Magyarországon az ezredfordulón*. JATE-Pressz, Szeged.

Szabó, G. (2006): Anchors aweigh! An analysis of the impact of anchor item's number and difficulty range on item difficulty calibrations. In: Nikolov, M. és Horváth, J. (szerk): *UPRT 2006: Empirical studies in English applied linguistics*. Lingua Franca Csoport, Pécs. 249–262.

Verhelst, N. (2004): Az item-válasz elmélet (IRT). In: Banerjee, J., Kaftandjieva, F., Takala, S. és Verhelst, N.: *Szintillesztési módszertani segédlet. A nyelvvizsgák illesztése a Közös európai referenciakerethez című kézikönyv előzetes, kísérleti verziójához.* Nyelvvizsgát Akkreditáló Testület – PH Nyelvvizsgáztatási Akkreditációs Központ, Budapest. 151–201.

Vígh Tibor (2005): A kommunikatív tesztelés elméleti alapjai. *Magyar Pedagógia*, **105**. 4. sz. 381–407.

Vígh Tibor (2007a): A kommunikatív tesztelés az idegen nyelvi mérés és értékelés rendszerében. Tematikus előadás. In: Korom Erzsébet (szerk.): *PÉK 2007 – V. Pedagógiai Értékelési Konferencia: Program – Tartalmi összefoglalók.* SZTE, Neveléstudományi Doktori Iskola, Szeged. 78. (http://www.staff.u-szeged.hu/~tvigh/Vigh_2007.pdf)

Vígh, T. (2007b): Bewertung von Schreibfertigkeit im Abitur für Deutsch als Fremdsprache. *Deutschunterricht für Ungarn*, 1–2. sz. 81–95.

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

ABSTRACT

TIBOR VÍGH: METHODOLOGICAL QUESTIONS OF BUILDING AN IRT-BASED ITEM BANK FOR FOREIGN LANGUAGE SKILLS

Results of analyzing of the Hungarian Matura exam tasks in German language

Launched in Hungary in 2005, the new foreign language Matura exam aims to assess and evaluate communicative language competence, conforming to the Common European Framework of Reference (CEFR). Though the ordinary and the advanced levels of the exam are designed to correspond to levels A2-B1 and B2 CEFR levels, respectively, there has been no empirical evidence to confirm this link. This paper proposes (1) to examine the extent to which the empirical difficulty levels of the Matura tasks cover the ability levels of exam-takers; (2) to investigate the possibilities of creating an IRT based language item bank representing the CEFR levels; and (3) to investigate the methodological issues and problems emerging in this process. The first part gives an overview of the bases and models of item response theory (IRT), as well as its application for assessing language skills. The second part discusses the necessity and problems of the empirical foundation of Matura language levels. The third part presents the results of the analysis of a data base. The 2006 ordinary level German as a foreign language Matura items in reading, listening comprehension and use of German were examined using the Rasch model. The fourth part presents the independent analysis of the person-item maps of the 2005 and 2006 ordinary and advanced level German Matura booklets. Arguments are presented for the necessity of examining the equivalence of booklets, for the analysis of the two Matura levels and the CEFR levels in a new assessment with the use of anchor and reference items, and for the creation of an IRT based item bank. Findings from European projects (e.g. Dialang, Dutch grid and EBAFLS) may be useful in selecting reference tasks and in characterising texts and items for the Matura tasks. The proposed item bank could function as a reference point for the development of exam tasks and could contribute to the empirical verification of the correspondence between the Matura and the CEFR language levels.

Magyar Pedagógia, **108**. Number 1. 29–51. (2008)

Levelezési cím / Address for correspondence: Szegedi Tudományegyetem, Neveléstudományi Doktori Iskola, H–6722 Szeged, Petőfi S. sgt. 30–34.