Pécsi Tudományegyetem
Bölcsészettudományi Kar
Nyelvtudományi Doktori Iskola
Alkalmazott Nyelvészeti Program

Werner Réka

A MÉLTÁNYOSSÁG NYOMÁBAN

A nemek szempontjából eltérő itemműködés (DIF) vizsgálata német nyelvi ECL beszéd- és szövegértés feladatokban

Doktori (PhD) értekezés tézisei

Témavezető:

dr. habil. Szabó Gábor PhD

Pécs

2023

Tartalomjegyzék

1. Bevezetés 2
2. Szakirodalmi áttekintés 3
2.1. A nyelvtudásmérő tesztek jóságmutatói 3
2.2. Validitás 5
2.3. Méltányosság (fairness) 8
2.4. Eltérő itemműködés (DIF) és itemtorzítás 10
3. Az ECL vizsgarendszer 13
4. Kutatási célok és kutatási kérdések 14
5. A vizsgálat folyamatának és korlátainak ismertetése 15
6. A kutatás összefoglalása és fő eredményei 18
7. Tanulságok és következtetések 24
8. Kitekintés 28
Irodalomjegyzék 30
Az értekezés témakörében releváns saját publikációk 34

1. Bevezetés

Magas kockázatú vizsgák – mint például a nyelvvizsga – esetén különösen fontos, hogy a mérés során felhasznált feladatok valóban a mérni kívánt konstruktumot mérjék, és a vizsgaeredmények ne legyenek összefüggésben egyéb, konstruktum-irreleváns tényezővel. Így például a vizsgázók neme, életkora, anyanyelve, származása nem befolyásolhatja a vizsga eredményét, illetve nem okozhatja az esetleges sikertelenséget.

A nyelvvizsgaközpontok egyik fő célkitűzése, hogy egyenlő esélyt biztosítsanak a vizsgázóknak nemüktől, koruktól, anyanyelvüktől és származásuktól függetlenül. Ennek érdekében a feladatírók már a feladatírás szakaszában igyekeznek elkerülni azokat az itemeket, amelyek feltehetően előnyben részesítik, vagy hátrányba helyezik a különböző vizsgázói alcsoportokat. Az ilyen jellegű itemek kiszűrése a tesztfejlesztési szakaszban kvalitatív módszerek alkalmazásával történik. Azonban a kvalitatív eljárások hatékonyságának ellenőrzése érdekében szükség van statisztikai módszerek alkalmazására is. Az értekezés célja a kvalitatív eljárások hatékonyságának ellenőrzése statisztikai módszerek segítségével. Az elemzés során a beszéd- és szövegértés mérésére használt itemek működését a klasszikus és a probabilisztikus tesztelmélet segítségével vizsgálom. A klasszikus elemzés lehetővé teszi a feladatok megbízhatóságának, az itemek nehézségi fokának és diszkriminációs indexének összehasonlítását. A modern, probabilisztikus tesztelmélet segítségével pedig az itemek modellszerű működése ellenőrizhető.

Jelen értekezés a 2018 és 2019-es években ECL nyelvvizsgán felhasznált német nyelvű B2 szintű feladatsorokra koncentrál. A vizsgálat teljeskörű, hiszen a megnevezett időszakban a teljes vizsgázói létszám itemszintű eredményeit feldolgozza.

Az értekezés célja egyrészt annak felderítése, hogy az ECL beszéd- és szövegértés vizsgafeladatok tartalmaznak-e olyan itemeket, amelyek megoldásában szignifikáns különbség mutatkozik a különböző nemű vizsgázók között, és ezáltal hátrányba vagy előnybe helyezik valamelyik vizsgázói csoportot. Másrészt az értekezés az eltérő itemműködés okait kívánja feltárni azáltal, hogy vizsgálja az adatokban található esetleges rendszeres mintázatokat, illetve felderíti, hogy a vizsgált itemek valóban előnyben vagy hátrányban részesítik-e a vizsgált nemi csoportok egyikét, ezáltal torzítva az eredményeket.

2. Szakirodalmi áttekintés

2.1. A nyelvtudásmérő tesztek jóságmutatói

Bachman (1990) a megbízhatóságot és az érvényességet tekinti a nyelvi értékelés két alapkövetelményének. A megbízhatóság (reliabilitás) a mérés pontosságára vonatkozik, és a tesztnek arra a tulajdonságára, amely az elméletben lehetővé teszi, hogy amennyiben a vizsgázók ugyanazt a tesztet, ugyanolyan körülmények között megismétlik, azonos vagy közel azonos eredményeket érjenek el (Bachman, 1990: 160-161). „A reliabilitás tehát magának a mérőeszköznek a megbízhatósága, amely más körülmények és populáció esetén is nagyjából eredeti tulajdonságait megtartva működik" (Bárdos, 2002: 38).

Mivel a nyelvtudás absztrakt és nem közvetlenül megfigyelhető, csak a mérés során elért eredmény alapján tudunk rá következtetni. A mérés megbízhatóságának kiszámításakor egy olyan modellre van szükség, amely arra enged következtetni, hogy milyen kapcsolat van a mérni kívánt képesség és a valós pontszám között (Bachman, 1990, 166). Ezt a célt szolgálja a Spearman által meghatározott valós pontszám-modell, melynek alapja az az elképzelés, hogy minden egyénileg elért pontszám (a megfigyelt pontszám) két fő összetevőből áll: a valós pontszámból és a hibából (Crocker & Algina, 1986: 107).

A megfigyelt pontszám tehát a nyelvvizsgák esetén a teszten elért eredmény, amelyet a valós pontszám és a hiba együttesen határoz meg. A valós pontszám úgy határozható meg, mint az az eredmény, amelyet egy személy a mérés során elérne, ha minden körülmény ideális lenne, és a mérési folyamat nem tartalmazna hibákat. A hiba pedig a megfigyelt pontszám és a valós érték közötti különbséget jelenti. A hibát eredményezheti a tesztfejlesztésből, a tesztelési körülményekből, vagy más tényezőkből eredendő hiányosság, pontatlanság (Bachman, 1990: 167, Bárdos, 2002: 38).

Bachman (1990: 164-165) azokat a tényezőket, amelyek hibát okozhatnak, és ezáltal kihathatnak az eredményre és a mérés megbízhatóságára három kategóriába sorolja: (1) módszereffektus, (2) a vizsgázók azon tulajdonságai, melyek nincsenek összefüggésben a mérni kívánt konstruktummal és (3) a véletlenszerű tényezők. Az első kategóriába sorolhatóak például a teszt hossza, az itemek/feladatok minősége, az alkalmazott feladattípus, a tesztadminisztráció és az értékelők megbízhatósága. A második kategóriába a vizsgázók neme, kora, származása és háttértudása sorolható, a harmadikba pedig olyan véletlenszerű tényezők, mint például a vizsgázók fáradtsága, vagy a találgatás, amelyek a pontszámok

következetlenségéhez vezethetnek, és befolyásolhatják a megbízhatóságot (Bachman, 1990: 164).

A megbízhatóságra tehát számos olyan tényező is hatással van, amelyek véletlenszerűek, nem ellenőrizhetők. Mivel ezeket sem előre megjósolni, sem kontrollálni nem lehet, Balchman és Palmer (1996: 20-21) arra hívják fel a figyelmet, hogy a teszt tervezésekor törekedni kell a kontrollálható tényezők csökkentésére, amelyek például a teszt formátumából, a teszt lebonyolításából és az értékelésből adódhatnak, és veszélyeztetik a megbízhatóságot.

A klasszikus valós pontszám-modell kapcsán aggályok merültek fel, ugyanis a különböző hibafajták nem elszigetelten jelennek meg a mérés során, a hibák közötti kölcsönhatások vizsgálata pedig nem lehetséges. Emellett a modell nem tesz különbséget a véletlenszerű és rendszerszerű hibák között, így nem teszi lehetővé a rendszerszintű hibák felismerését és orvoslását (Bachman, 1990: 185-187). Ezek a hiányosságok nyilvánvalóvá tették, hogy a megbízhatóság kiszámításához rugalmasabb, komplexebb szempontokat figyelembe vevő eljárásokra van szükség. Ilyen reliabilitás számítási módszerek például: az ismétlés (test-retest), a párhuzamos tesztelési (parallel forms method) és az úgynevezett felezős módszerek (split-half method) (Bárdos, 2002: 39; Crocker & Algina, 1986: 132-134; Bachman, 1990: 169-177).

A megbízhatóság mellett Bachman (1990) a tesztek legfontosabb jóságmutatójának az érvényességet tekinti. Az érvényesség (validitás) alapvetően úgy határozható meg, hogy a teszt azt méri, „amit eredetileg mérni szándékozott” (Bárdos, 2002:40). Ahhoz azonban, hogy ezt vizsgálni tudjuk, szükség van annak pontos meghatározására, hogy mi a mérés célja, mire használhatóak a vizsga eredményei, milyen következtetések vonhatóak le az eredmény alapján (Bachman, 1990: 236).

A megbízhatóság és az érvényesség nem vizsgálható egymástól függetlenül, hiszen megbízhatóság nélkül egy teszt nem tekinthető érvényesnek. Egymagában a megbízhatóság viszont még nem garantálja a teszt és a tesztelési folyamat érvényességét. A megbízhatóság tehát előfeltétele a validitásnak. Tágabb értelemben véve a megbízhatóság hozzájárul a mérés általános érvényességéhez, mivel biztosítja, hogy a mérés és értékelés következetes és megbízható legyen. Azonban a validitás továbbmegy a megbízhatóságon, mivel azt is vizsgálja, hogy az eszköz pontosan méri-e a kívánt konstruktumot, és helyes következtetésekhez vezet-e (Bárdos, 2002; Elder, 2012; Bachman, 1990: 236). Ennek az összefüggésnek köszönhetően a szakirodalomban a későbbiekben a megbízhatóságot nem önálló jóságmutatóként emlegetik, hanem a validitás részeként tüntetik fel (Bachman & Palmer, 1996; Vígh, 2005; ALTE, 2020).

Bachman és Palmer (1996) a teszt hasznosságát (usefulness) tekintették a legfontosabb kritériumnak. A hasznosságot a konstruktum validitás (construct validity), megbízhatóság (reliability), autenticitás (authenticity), interaktivitás (interactivness), hatás (impact) és praktikusság (practicality) ötvözeteként fogalmazták meg. Az alapelveik szerint a mérések során nem csak az egyes tulajdonságokat kell javítani, hanem törekedni kell a hasznosságot összességében maximalizálni.

A nemzetközi mérési és nyelvvizsgáztatási szakirodalom a tesztek jóságmutatóiként a tartalmi érvényességet, a konstruktum és kritériumorientált érvényességet, valamint a megbízhatóságot említik (AERA et al., 1999; 2014). Az Association of Language Testers in Europe[1] (továbbiakban ALTE) összesen nyolc (ALTE, 2020) minőségi szempontot sorol fel: tartalmi érvényesség, konstruktum érvényesség, megbízhatóság, kritériumorientált validitás, méltányosság, a szolgáltatás minősége, praktikusság és teszthatás, melyek segítségével érhető el a tesztelés validitása. Egy vizsgát akkor tekintenek érvényesnek, ha ezek a szempontok együttesen valósulnak meg a teljes vizsgafolyamat során. Ezek a minőségi szempontok túlmutatnak a tesztfejlesztési szakaszon, hiszen kiterjednek többek között a teljes vizsgaadminisztrációra, a vizsgázók tájékoztatására és az eredmények értelmezésére is.

Összegzésként megállapítható, hogy a Bachman és Palmer által meghatározott jóságmutatók, valamint a nyelvtudásmérő szakmai egyesületek (például ALTE) által megnevezett jóságmutatók között számos közös elem található, azonban az aktuális szakirodalom további specifikus jóságmutatókat is említ, így például a méltányosságot (AERA et al., 1999; 2014; ALTE, 2020).

2.2. Validitás

A validitás a nyelvvizsgák kontextusában alapvetően azt jelenti, hogy egy vizsga valóban azt méri, amit célul tűzött ki (AERA et al., 2014: 11; Bachman, 1990: 25; Wesolowski & Wind, 2019: 438). A teszttel mérni kívánt hipotetikus képességet a szakirodalom konstruktumnak nevezi (Bachman, 1990: 43). A nyelvi tesztelésben ez általában egy adott nyelvi készségre utal, például a hallás, olvasás, írás vagy beszéd készségére (ALTE, 1998: 139). A validitás a

[1] Az Association of Language Testers in Europe (ALTE) egy olyan európai szervezet, amelynek egyik fő célja a nyelvi tesztelés minőségének, megbízhatóságának és érvényességének biztosítása. A szervezet tagjainak rendelkezniük kell egy jól kidolgozott és megalapozott minőségellenőrzési rendszerrel, amely biztosítja a nyelvi tesztek megbízhatóságát és érvényességét. Ezt az ALTE egy auditálási folyamat során ellenőrzi, és amennyiben az auditált nyelvvizsgarendszer minőségbiztosítási elvei megfelelőnek bizonyulnak, az adott vizsga elnyeri az ALTE minőségi védjegyét (*Q-Mark*).

nyelvvizsga esetén tehát azt jelzi, hogy egy vizsga képes-e megbízhatóan és hatékonyan mérni és értékelni a vizsgázó tudását, készségeit vagy kompetenciáit az adott nyelvi területen.

A validitás fogalmának számos egyéb értelmezése létezik. Cronbach és Meehl a *Construct validity in psychological tests* (1955: 175) című alapműben négyféle érvényességtípust határoz meg: a prediktív érvényességet (predictive validity), a konkurens érvényességet (concurrent validity), a tartalmi érvényességet (content validity) és a konstruktum validitást (construct validity). Samuel Messick (1989) átfogó módon határozta meg az érvényességet, és két fő típusát különbözteti meg: a tartalmi érvényességet és a kritériumorientált érvényességet. Messick szerint az érvényesség egy többdimenziós fogalom, amely a teszteredmények értelmezéséhez és felhasználásához kapcsolódik. Definíciója hangsúlyozza, hogy a teszt érvényessége nem tekinthető fix tulajdonságnak, hanem keretként szolgál a teszteredmények értelmezéséhez és felhasználásához (Messick, 1989: 13). Messick elmélete szerint egy teszt akkor tekinthető érvényesnek, ha a vizsgázók elért eredményei lehetővé teszik a döntéshozók és felhasználók számára, hogy érvényes következtetéseket vonjanak le a teszt eredményei alapján.

Fontos hangsúlyozni, hogy a Messick-féle megközelítés a vizsga teljeskörű validitását jelenti, mivel az érvényesség mindig csak a vizsga valamelyik aspektusára vonatkozik, például a tesztfejlesztésre vagy az értékelésre. Így a vizsgaközpontok feladata, a vizsga teljes folyamatát átívelően, olyan eljárási módok kidolgozása, amelyek segítségével megvalósul az érvényesség.

A validitásnak tehát többféle megközelítési módja létezik, de alapvetően a Cronbach és Meehl (1955) által meghatározott három kategóriát említi a szakirodalom: a tartalmi érvényességet, a kritériumorientált érvényességet és a konstruktum validitást (Messick, 1989; Bachman & Palmer, 1996; AERA et al., 1999; 2014; ALTE, 2020).

A felsorolt három validitástípus mellett a látszatvaliditás (*face validity*) fogalmával találkozunk még a szakirodalom (Alderson et al., 1995; Urbina, 2004), amely egy szakértők vagy akár laikusok által hozott ítélet arról, hogy a mérőeszköz azt méri-e, amit állít. A látszatvaliditás vizsgálata tehát nem statisztikai elemzéseken, hanem holisztikus jellegű megítélésen alapul.

Messick (1989) az érvényesség fogalmát kibővíti a *consequential validity*-vel, és kiemeli a következmények fontosságát, melyek az eredmény felhasználásából adódnak. Ennek relevanciáját McNamara (2000: 53-54) és Kane (2010: 179) is felismerik, és rávilágítanak arra, hogy a validitás szempontjából a mérés társadalmi és etikai következményeit már a mérés tervezésekor is figyelembe kell venni, illetve annak felhasználásakor is. Ez különösen fontos a magas kockázatú vizsgák esetén, ahol a vizsga kimenetele jelentős következményekkel járhat az

egyén, az oktatási intézmények és politikai döntések szempontjából is (Bachman and Palmer 1996: 279-285; Kane, 2011: 12). A *consequential validity* fogalma kiemeli a vizsgafejlesztők etikai felelősségét és annak szükségszerűségét, hogy már a tesztfejlesztés szakaszában figyelembe vegyék és minimalizálják a negatív következményeket, miközben maximalizálják a mérés pozitív hatását (McNamara, 2000).

A megbízhatósággal ellentétben a validitás nem matematikai módszerekkel kiszámítható tulajdonság, hanem megítélés kérdése (McNamara & Roever, 2006: 10). A validálás során ugyanis nem a tesztet magát kell érvényesíteni, hanem a teszteredmény alapján levont következtetések érvényességét kell vizsgálni (Cronbach & Meehl, 1955: 297). A validálás tehát egy komplex folyamat, amely döntések sorozatából áll, és jól kidolgozott keretrendszerek segítségével valósítható meg. Ezek a keretrendszerek iránymutatást nyújthatnak a mérési eszközök fejlesztéséhez, a vizsgafolyamat lebonyolításához és az eredmények felhasználásához (Kane, 2011: 8).

Ebben az összefüggésben Mislevy egy bizonyítékalapú tesztfejlesztési keretrendszert (*Evidence-Centered Design*) dolgozott ki, mely Messick validitás értelmezésén alapul. Keretrendszere a tesztfejlesztési folyamatot helyezi középpontba, és nem fordít figyelmet az eredmények alapján levont következtetésekre (Mislevy et al., 2002, 2003). Mislevy értelmezésétől eltérően Kane (2013: 3) kiemeli, hogy a „validitás nem a teszt tulajdonsága", hanem inkább azokban az értelmezésekben és felhasználásokban rejlik, amelyek a tesztpontszámokon alapulnak. Az érvényesség olvasatában azt jelenti, hogy az eredmények értelmezése és felhasználása ésszerű és megfelelő bizonyítékokkal alátámasztott.

Bachman és Palmer az *Assessment Use Argument* elnevezésű keretrendszerük kidolgozásakor a pontszámok értelmezésének és alkalmazásának validálását helyezték a középpontba. A keretrendszer hangsúlyozza, hogy a validálás strukturált következtetések sorozatából áll, célja pedig a nyelvi értékelés különböző aspektusainak – például a tesztfejlesztés folyamata és az eredmények következményeinek – az összehangolása (Bachman & Palmer, 2010: 30).

Kane nevéhez is fűződik egy validálási keretrendszer. Mivel Kane (2011: 8) a hagyományos konstruktum érvényességi megközelítést (Bachman & Palmer, 1996) meglehetősen absztrakt folyamatnak tekintette, ami nem biztosít kellő iránymutatást a gyakorlatban történő alkalmazáshoz, egy gyakorlatiasabb és strukturáltabb megközelítés szükségességét fogalmazta meg, és egy érvelésalapú validációs keretrendszert dolgozott ki. Kane hangsúlyozza, hogy a validálás nem csupán annak ellenőrzése, hogy egy teszt általánosságban

azt méri-e, amit hivatott, hanem a validálás folyamata során azokat a konkrét következtetéseket és döntéseket kell kidolgozni és megalapozni, amelyeket a teszteredmények alapján hoznak majd (Kane, 2011).

2.3. Méltányosság (fairness)

A méltányosság fontosságát már Cronbach is kiemeli „Essentials of Psychological Testing" című művében. A szerző hangsúlyozza, hogy egy tesztnek méltányosnak és elfogulatlannak kell lennie, annak érdekében, hogy minden egyén számára egyenlő esélyt biztosítson tudása és képességei bizonyítására. Cronbach megfogalmazza, hogy a torzító elemeket el kell kerülni a tesztelés során, és biztosítani kell, hogy a tesztek és a tesztadminisztráció eljárásai ne részesítsenek előnyben vagy hátrányban egyetlen vizsgázói csoportot sem (Cronbach L. J., 1984).

Messick rávilágított arra, hogy a tesztméltányosság a konstruktum validitás elengedhetetlen eleme, mivel egy teszt nem tekinthető érvényesnek, ha nem méltányos minden résztvevő számára (Messick, 1989). A méltányosság fogalma az utóbbi évtizedekben egyre gyakrabban jelenik meg a nyelvi értékeléssel kapcsolatos írásokban, tanulmányokban és előadásokban is (Bachman & Palmer, 2010; Kane, 2010; Kremmel, 2019; Kunnan, 2000; 2004; 2007; 2014; Stoynoff, 2012). Szakmai kiadványok, mint például a *Code of fair testing practices in education* (Joint Committee on Testing Practices, 2005: 23), *ETS International Principles for the Fairness of Assessments* (ETS, 2016: 3-4) és az *ALTE Principles of Good Practice* (ALTE, 2020: 13) is hangsúlyozzák, hogy a tesztek fejlesztőinek törekedniük kell arra, hogy a tesztjeik a lehető leginkább igazságosak legyenek a különböző nemű, életkorú, etnikai hovatartozású, kulturális és nyelvi háttérrel rendelkező, hátrányos helyzetű és különleges szükségletekkel rendelkező vizsgázók számára. Ennek megfelelően a teszteknek biztosítaniuk kell, hogy az összes vizsgázónak lehetősége legyen megmutatni tudását az adott nyelvi készségen, és a feladatok ne részesítsék előnyben vagy hátrányban az egyes résztevevőket konstruktum-irreleváns tényezők miatt (AERA et al., 2014: 50; Bachman & Palmer, 2010: 128-129). Ennek a célkitűzésnek az eléréséhez az egyik leghatékonyabb stratégia a torzításmentes tesztek készítése. A tesztfejlesztőknek tehát biztosítaniuk kell, hogy a feladatok ne tartalmazzanak olyan tényezőket, amelyek nem kapcsolódnak a mérendő területhez, és ezáltal előnyt vagy hátrányt jelentenek valamelyik vizsgázói csoport részére (ALTE, 1998: 136; Bachman & Palmer, 2010: 128).

A tesztelés méltányosságát a teljes tesztelési folyamat során biztosítani kell. A tesztadminisztráció során ezt standardizált utasítások, hasonló vizsgakörülmények biztosításával, valamint teremfelügyelői útmutatók kidolgozásával és betartatásával lehetséges elérni. A méltányosság biztosítása érdekében azonban nem kizárt olyan formai módosítások végrehajtása, amelyek lehetőséget teremtenek arra, hogy például más képességű vizsgázók is részt vehessenek a vizsgán. A módosítás kizárólag formai jellegű lehet, a tartalmat, a mérni kívánt konstruktumot nem érintheti (Bachman & Palmer, 2010: 128; AERA et al., 2014).

Ugyanakkor a méltányosság nemcsak a résztvevők egyenlő bánásmódjának biztosítása miatt fontos, hanem a pontos, megbízható teszteredmények végett is. A *Standards for Psychological and Educational Testing* szerkesztői arra is felhívják a figyelmet, hogy a méltányosságra nemcsak a teszt maga hat, hanem számos egyéb tényező is, mint például az eredmények közlésének és értelmezésének módja, az eredmény felhasználásának következményei. A méltányosságot abban az esetben vizsgálnánk átfogóan, ha azokat a rendeleteket és törvényeket is vizsgálnánk, amelyek a tesztek – a nyelvvizsgák esetén az eredmények – felhasználást szabályozzák (AERA et al., 2014: 59). Ez különösen fontos felvetés a nyelvtudást igazoló vizsgák esetében, mivel az eredmények döntő hatással lehetnek az egyén továbbtanulási vagy munkavállalási lehetőségeire, valamint egy adott országba történő bevándorlásra vonatkozó követelmények teljesítésére. Az ECL német mint idegen nyelv vizsga is a magas kockázatú vizsgák közé sorolható, hiszen többek között sikeres vizsga esetén a Magyarországon egyetemi tanulmányokat folytató vizsgázók pluszpontokat szereznek a felvételi eljárás során (2012. évi 423. (XII. 29.) Korm. rendelet), a külföldi vizsgázók pedig meghatározott szintű ECL nyelvvizsgabizonyítvány tulajdonában németországi bevándorlásukat kezdeményezhetik. A Német Külügyminisztérium azon vizsgák listáján jegyzi a német nyelvű vizsgát, amelyek megléte igazolja a bevándorláshoz vagy munkavállaláshoz szükséges általános nyelvi ismeretek meglétét (Auswärtiges Amt, 2023: 450).

A nyelvi tesztelésben számos tényező van jelen, melyek a méltányosságot veszélyeztetik. Ezek a tényezők a *Standards for Psychological and Educational Testing* megfogalmazásában vagy a tesztelés folyamatában vagy magában a tesztben vannak jelen. A szerkesztők négy tényezőt határoznak meg, melyek veszélyeztetik a méltányosságot: (1) teszttartalom, (2) tesztkontextus, (3) tesztválasz és (4) tanulási lehetőség (AERA et al., 2014: 54-57). A méltányosság biztosítása érdekében a tesztfejlesztőknek oda kell figyelniük ezekre a tényezőkre, és megfelelő módon kezelniük azokat a tesztírás és a tesztfejlesztés során. A teljes vizsgafolyamat méltányosságát szem előtt tartva biztosítaniuk kell, hogy a nyelvi tesztek valóban

igazságosan és objektíven mérjék a vizsgázók nyelvi képességeit (Bachman & Palmer, 2010: 128).

2.4. Eltérő itemműködés (DIF) és itemtorzítás

Ahogy Kunnan (2004: 33) is hangsúlyozza, a tesztfejlesztőknek erőfeszítéseket kell tenniük a méltányos, azaz torzításmentes tesztek készítése érdekében. Shepard (1982: 14) a torzítást „egy csoport teljesítményét hátrányosan érintő szisztematikus hibaként" határozza meg. Ennek megfelelően akkor beszélünk item- vagy feladattorzulásról, ha az itemen vagy a feladaton elért eredmények különbsége nem a mért készséget tükrözi, hanem a tesztelők olyan egyéni jellemzőiben rejlik, amelyek nem kapcsolódnak a konstruktumhoz (például a vizsgázók neme, életkora, etnikai hovatartozása, szakmai tapasztalata stb.) (Bachman, 1990: 271; Camilli & Shepard, 1994: 8; Bachman & Palmer, 2010: 129). Torzításról beszélünk például akkor, ha egy beszédértés feladat a férfiak számára könnyebb, mint a nők számára, pusztán a nemekre jellemző tapasztalatok, érdeklődés és háttérismeret miatt, nem pedig a hallási képességük miatt. Mivel a beszédértés teszt csak a tesztelők halláskészségét kívánja mérni, a kérdéses item torzít azáltal, hogy előnybe részesíti az egyik csoportot, és ezzel egyidőben hátrányba hozza a másikat.

A validitáshoz és a méltányossághoz hasonlóan a torzítás is meglehetősen összetett, számos tényező okozhatja, mint például a tesztpontszámok helytelen értelmezése, bizonyos csoportokat előnybe helyező teszttartalom, rosszul megválasztott feladattípus vagy méltánytalan körülmények. A torzítás feltárása viszonylag nehéz, ugyanis a mérni kívánt konstruktumot nehéz egyértelműen leválasztani a vizsgázó tapasztalataitól, a kulturális és az oktatási háttértől (Bachman, 1990: 272; Bachman & Palmer, 2010: 128). „Az értékelési szakemberek és oktatók kihívása abban rejlik, hogy feltárják és megértsék, hogy hogyan befolyásolják a kulturális és nyelvi tényezők az értékelést ahelyett, hogy tagadnák, hogy ilyen hatások könnyedén megjelenhetnek" (Duran, 1989: 573-574; idézi Bachman, 1990: 272).

Bachman (1990:271) szerint a torzítás oka mind a tesztírásban, mind a felhasználásban keresendő. Annak érdekében, hogy elkerüljék vagy legalább csökkentsék a feladat- és teszttorzulást, a tesztfejlesztőknek már a tesztek írásakor tisztában kell lenniük a torzulás lehetséges forrásaival. A nyelvi tesztelés területén végzett tanulmányok már azonosítottak különféle torzítást okozó tényezőt, ideértve az életkort, a nemet, a kulturális és etnikai hátteret, az anyanyelvi hátteret, a konkrét szakterülethez kapcsolódó háttérismereteket és tapasztalatokat stb. (Bachman, 1990: 271-279; Elder, 2012: 1; Kunnan, 2007: 110). Mivel azonban a feladatfejlesztésben részt vevő szakemberek nem minden esetben ismerik fel a torzítás lehetséges

forrásait, kérdéses marad, hogy a tesztfejlesztés során alkalmazott minőségi eljárások elegendőek-e a tesztek méltányosságának biztosításához. Ezért az éles vizsga után további elemzésekre van szükség. Ennek egyik módja az itemek eltérő működésének (DIF) vizsgálata (AERA et al., 2014: 51; ALTE, 2011: 80; Elder, 2012: 2; ETS, 2016: 20; Ferne & Rupp, 2007: 114; Kunnan, 2007: 109).

A DIF-elemzés segítségével feltárt eltérő működést mutató itemek kizárása a tesztelési folyamatból a megbízhatóság növelésének egyik alapkritériuma (Kunnan 2004, 2008; McNamara, & Roever, 2006; AERA et al., 2014). A DIF jelenléte ugyanis mérési hibát okozhat, ami csökkenti az értékelés megbízhatóságát. A DIF azonosításával és megszüntetésével a vizsgarendszer tehát növelheti a teszteredmények és az eredmények alapján levont következtetések megbízhatóságát is.

A DIF feltárására több IRT-alapú statisztikai módszer létezik (Ross & Okabe, 2006). A megfelelő módszer kiválasztása számos tényezőtől függ, többek között az itemek típusától (dichotóm vagy nem dichotóm itemek), az összehasonlítandó csoportok számától, a minta méretétől és a konkrét kutatási kérdéstől. Az IRT modellek közös tulajdonsága, hogy megbecsülik az item paramétereket, és megvizsgálják, hogy azok hogyan változnak a különböző csoportok között. Az egyes itemekre adott helyes válasz valószínűségét nem a teszten elért összpontszámmal vetik össze, hanem ún. látensvonás-értékkel.

A DIF-elemzés ugyan lehetővé teszi a potenciálisan torzító (bias) itemek beazonosítását, azonban mindenképpen szükség van az érintett itemek további kvalitatív elemzésére is. Az eltérő itemműködés statisztikai kimutatása ugyanis nem mindig jelenti automatikusan azt, hogy az item valóban torzít (bias). A kvalitatív elemzések célja annak vizsgálata, hogy a kérdéses item mely jellemzői okozhatták az eltérő itemműködést (Ferne & Rupp, 2007: 141; Bachman & Palmer, 2010: 130).

Habár a DIF-elemzést hosszú ideje alkalmazzák oktatási területen, a nyelvi értékelés területen csak az elmúlt évtizedekben vált egyre elterjedtebbé. A nemek közötti különbségeket több kutatás vizsgálja, ezek alapvetően a nyelvi készségekre, a teszt tartalmára és formátuma fokuszálnak (Hyde & Linn, 1988; Chen & Henning, 1985; Cole, 1997; Zwick et al., 2004; Pae, 2004; Elder, 2012).

Annak ellenére, hogy a DIF-elemzés jelentősége kiemelkedő a nyelvvizsgák minőségének javításában, az ezzel kapcsolatos hazai kutatások hiánya (Kiszely, 2022), arra enged következtetni, hogy a rutinszerű alkalmazása a magyarországi vizsgaközpontok

gyakorlatába még nem épült be. A nemzetközi vizsgarendszerek azonban alkalmazzák a DIF-elemzést. Erről tanúskodnak a publikált kutatások, melyek különböző feltárási módszerek segítségével vizsgálják az eltérő itemműködés jelenlétét a vizsgákon alkalmazott feladatokban (Ryan & Bachman, 1992; Eckes, 2011; Song, 2014; Geranpayeh & Kunnan, 2007; Coniam & Lee, 2021).

Magyar viszonylatban a kutatásokban a DIF-elemzés nem jelenik meg a nyelvvizsgákkal kapcsolatosan. Azonban a férfiak és a nők sikerességét közép- és felsőoktatási kontextusban, valamint a nyelvvizsgán több kutatás is vizsgálta (Fényes, 2008; 2009; 2010a; 2010b). Kispál és Gergely (2020) DIF-elemzés segítségével, kétparaméteres logisztikus modellt alkalmazva a halmozottan hátrányos helyzetű diákok Országos kompetenciamérésen nyújtott teljesítményét vizsgálta.

Nemzetközi viszonylatban megállapítható, hogy a hosszú távú DIF-elemzés hiányzik a nyelvi értékelés területén (Pae, 2012: 534). Ebből kifolyólag az értekezés célja nem egyetlen időpontban (azaz egy vizsgaidőszakban), hanem több adatgyűjtési ponton át vizsgálni a nemek szempontjából eltérő itemműködést jelző itemek jelenlétét és az itemek által okozott esetleges torzítás okait. Magyar viszonylatban pedig Kiszely (2022) – a magyar folyóiratokban megjelent kutatások összefoglalójában – utal arra, „hogy Fekete és Csépes (2018) alapján további kutatásokat lehetne folytatni, amelyek a vizsgázói teljesítménykülönbségeket elemeznék nemek, életkorok, földrajzi elhelyezkedés és egyéb szempontok tükrében.” (Kiszely, 2022: 37). Jelen értekezés ugyan nem tér ki az összes megnevezett változóra, de kísérletet tesz a nemek szerinti különbségek feltárására és okainak vizsgálatára.

3. Az ECL vizsgarendszer

Az ECL nyelvvizsga a Pécsi Tudományegyetem Idegen Nyelvi Központja által működtetett egynyelvű, papír alapú, a kommunikatív nyelvi kompetencia elméleti modelljén alapuló vizsga, amelynek célja az általános nyelvtudás mérése. Az ECL tesztek a négy fő nyelvi készséget (beszédértés, szövegértés, íráskészség és beszédkészség) valós kommunikációs helyzetekben a KER négy szintjén (A2, B1, B2 és C1) mérik (Council of Europe, 2001: 14). Az szövegértést és az íráskészséget a vizsga írásbeli részében, míg a beszédértést és a beszédkészséget a szóbeli részében vizsgálják.

Az ECL vizsgarendszer a minőségbiztosítás keretében kiemelt figyelmet fordít a megbízhatóságra, a vizsga érvényességére és a méltányosságra, a tesztfejlesztés és a vizsgaadminisztráció minden szakaszában. A lehetséges torzítások felderítése már a tesztírás során elkezdődik, az alkalmazott módszerek hatékonyságát pedig az éles vizsgát követő kvantitatív elemzések ellenőrzik.

4. Kutatási célok és kutatási kérdések

Az ECL vizsgarendszer különös figyelmet fordít arra, hogy a tesztfejlesztés különböző szakaszaiban minőségi és mennyiségi eljárások alkalmazásával csökkentse azon itemek számát, amelyek előnybe helyezik vagy hátrányba hozzák a vizsgázók egyes csoportjait. Ebből kifolyólag a következő nullhipotézis fogalmazható meg:

> H: Az ECL B2-es szintű német nyelvi beszéd- és szövegértés feladatokban az éles vizsgát megelőző és azt követő minőségbiztosítási eljárások sokfélesége miatt csak korlátozott számú, a nemek szempontjából eltérő működést mutató item található, az itemek jelenléte azonban nem okoz szignifikáns különbséget a férfi és női vizsgázók feladatszintű teljesítményében.

Az értekezés elsődleges célja a nullhipotézis tesztelése. A hipotézis igazolásához vagy megcáfolásához a következő kutatási kérdés megválaszolására van szükség:

> K1: Felfedezhető-e a beszéd- és szövegértés feladatokban eltérő itemműködés a vizsgázók neme szempontjából? Ha igen, milyen mértékben?

A nullhipotézis tesztelése, és ezáltal a nemek szempontjából eltérő működést mutató itemek számának meghatározása, az értekezés két fő céljának egyike. Azon tényezők, amelyek esetlegesen a nemek közötti különbségeket okozzák, empirikusan is vizsgálandók. Az értekezés második célja olyan esetleges, szisztematikus mintázatok feltárása (pl. a feladattípust, a feladat nehézségét, a feladat tartalmát és az item általános illeszkedési mutatóit illetően), amelyek az itemek nemek szempontjából eltérő működésével szorosan összefüggenek, és ezáltal a torzítás potenciális forrásai lehetnek.

Az értekezés második fő céljának elérése érdekében két további kutatási kérdés megválaszolására van szükség:

> K2: Felfedezhetőek-e olyan szisztematikus mintázatok, amelyek rendszeresen az egyik csoportot hozzák előnybe, és az eredmények helytelen értelmezéshez vezethetnek?
>
> K3: A DIF-ként jelölt itemek ténylegesen előnyben részesítik, vagy hátrányosan érintik a vizsgált csoportok egyikét, és ezáltal torzítják az eredményt?

Az elméleti kerettel összhangban, mely kiemeli a teszt validitásának és méltányosságának fontosságát, a K2 kérdés a teszt validitását vizsgálja, a K3 kérdés fókuszában pedig a méltányosság áll.

5. A vizsgálat folyamatának és korlátainak ismertetése

A két fő célkitűzéssel összhangban a tanulmány két részből áll:

(1) a nemek közötti DIF-elemzésből és

(2) a DIF lehetséges okainak és a torzítások forrásainak feltárásából.

Az értekezés fókuszában az ECL nyelvvizsgán a 2018 és 2019-es évben[2] felhasznált német nyelvű B2 szintű szöveg- és beszédértés vizsgafeladatsorai állnak. A kutatás első lépéseként a beszéd- és szövegértés feladatsorok DIF-elemzéseit készítettem el. Az elemzések eredményeinek szisztematikus áttekintése során a vizsgált időszakban 13 beszéd- és 18 szövegértés item esetén fedeztem fel eltérő itemműködést. Második lépésként a DIF-elemzés eredményeit együtt vizsgáltam a klasszikus elemzések (itemnehézség, diszkriminációs mutatók és megbízhatósági mutatók) eredményeivel, majd megvizsgáltam az itemek általános illeszkedési mutatóit is.

Az értekezés pilotkutatásaként az eltérő itemműködést jelző beszédértés itemeket 9 szakértő (gyakorló nyelvtanárok, vizsgáztatók és mérésben jártas szakemberek) véleményezte. Egyrészt egy 5 fokú Likert-skála segítségével minden item esetén nyilatkoztak arról, hogy a férfi és a női vizsgázókat előnyben részesíti-e vagy hátrányosan érinti-e az item, másrészt szöveges indoklással támasztották alá az itemek besorolását. A pilotkutatás eredményeinek elemzése során nyilvánvalóvá vált, hogy további kutatásra és elemzésekre van szükség annak érdekében, hogy megbízható eredmény szülessen (Hrisztova-Gotthardt & Werner, 2021).

Míg a pilotkutatás csak a beszédkészég itemeket vizsgálta, jelen kutatást kiterjesztettem a szövegértés itemekre is, azzal a céllal, hogy átfogóbb képet kapjak az itemek működéséről mindkét készség esetén. A pilotkutatás során validált kérdőívet mindkét készségre vonatkozóan tapasztalt nyelvtanárok (többségükben az ECL Vizsgaközpont német nyelvből akkreditált vizsgáztatói) és mérés-értékelésben jártas szakértők töltötték ki. A beszédértés kérdőívet 212 fő, a szövegértés kérdőívet pedig 215 fő töltötte ki. A tanárok és szakértők minden item esetén nyilatkoztak arról, hogy az adott item a férfi és a női vizsgázók esetén előnyben részesíti vagy hátrányosan érinti az adott nem képviselőit. A kitöltés során a következő besorolások között

[2] A kutatás 2020-as évben kezdődött, amikor a Covid19-világjárvány miatt egyrészt kimaradtak vizsgaidőszakok, másrészt a nyelvvizsgakötelezettség eltörlése a vizsgázói létszám drasztikus csökkenéséhez vezetett. Nyilvánvalóvá vált tehát, hogy a 2020-as év adatai nem lesznek alkalmasak egy nagymintás feltáró kutatás lebonyolításához. Annak érdekében, hogy az elemzést lehetőségekhez mérten magas létszámú populáción végezzem, a 2018 és 2019-es évek mellett döntöttem. A döntést indokolja az is, hogy az ECL Vizsgaközpont gyakorlatába beépült az az eljárás, miszerint azokat a feladatok, amelyek statisztikai mutatói megfelelőnek bizonyulnak, két-három év elteltével újra felhasználja. Így az elemzés célja annak felderítése, hogy a különböző nemek szempontjából hogy működtek a feladatok, van-e szükség módosításra, illetve az itemek átdolgozására egy esetleges újabb felhasználás előtt.

választhattak: (1) erős előny, (2) előny, (3) semleges, (4) hátrány, (5) erős hátrány. A besoroláson kívül minden item esetén szöveges indoklást írtak megítélésük alátámasztásaként. A kérdőív segítségével gyűjtött szöveges indoklásokat kvalitatívan elemeztem. Emellett az elemzés során egyrészt a skálaértékek átlagát, másrészt a gyakorisági eloszlást vettem figyelembe.

A kérdőíves kutatás kiegészítéseként interjúkra került sor, melyeken egyrészt 5 nyelvtanár, illetve mérési szakember, másrészt 9 diák vett részt. A résztvevők az interjúk során arról nyilatkoztak, hogy a vizsgált beszéd- és szövegértés itemek előnyösek-e vagy hátrányosak-e a férfi vagy női vizsgázók számára. Az interjúk résztvevői kitértek az elfogultság potenciális forrásaira, beleértve a tartalmat, a mért készségeket, a feladattípust, a szókincset, a szövegek témáját és a vizsgázók lehetséges tapasztalataira az adott témakörben.

A kérdőívek kiértékelése után kiválasztottam azokat a rövid válasz típusú itemeket, amelyek a szakértők véleménye szerint leginkább előnyhöz juttatták az egyik vizsgázói csoportot, és ezáltal hátrányt okoztak a másik csoportnak. Annak érdekében, hogy a vizsgázói válaszok kvalitatív elemzését is elvégezhessem, a vizsgázói válaszokat is rögzítettük.

A vizsgált időtartamot tekintve a kutatás teljes körű, az ECL Vizsgaközpont minden vizsgázójának az eredményét feldolgozza, aki 2018. február és 2019. december között vett részt német nyelvből a beszéd- vagy a szövegértés vizsgarészen. Mivel a vizsgázói létszám az adott időpontokban mindössze a B2 szinten haladta meg férfiak és a nők esetén is a 100 főt, a kutatás kizárólag a B2 szintű vizsgák eredményeit vizsgálja. Az egyparaméteres IRT modellek alkalmazásához – amelyek közé a *Facets* program is sorolható – ugyanis minimálisan 100 fő javasolt (Alderson et al., 1995: 91). A DIF-elemzésre vonatkozóan a *Facets* program kézikönyve nem tartalmaz konkrét javaslatot, mindössze egy utalást arra, hogy a lehető legnagyobb mintán kell elvégezni az elemzést (Clauser & Hambleton, 1994: 91). Ebből kifolyólag az általános javaslatot tekintettem mérvadónak. Mivel az A2, B1 és C1 szinten a női és a férfi vizsgázók száma egy esetben sem éri el a 100 főt, ezeket a szinteket az elemzésből kizártam.

Ugyan a DIF-elemzés alkalmas lenne az itemek kor szerinti működésének vizsgálatára is, ettől jelen értekezés eltekint. Az egyes korcsoportokhoz tartozó vizsgázók létszáma ugyanis sok esetben nem éri el a 100 főt, így az eredmények megbízhatósága megkérdőjelezhető lenne.

A vizsgált populáció összetétele a nemzetiséget tekintve homogén, a vizsgált időszakban a beszédkészség vizsgázók 98,01%-a, a szövegértés vizsgázók 98,29%-a magyar anyanyelvű, így az anyanyelv szempontjából a méltányosság vizsgálata nem releváns.

A kutatás kevert módszertanú (Creswell J. W., 2004; Dörnyei, 2007; Mackey & Gass, 2005), felépítéskor a Creswell és Plano Clark (2011: 83) által meghatározott tipológiarendszert követve a magyarázó, egymásra épülő felépítés modelljét alkalmaztam (Creswell & Plano Clark, 2011; Sántha, 2013; Király et al, 2014; Tashakorri & Teddlie, 2003). A kutatás két szakaszból állt. Az első, kvantitatív szakasz eredményeit és megállapításait használtam fel a második, kvalitatív szakasz kidolgozására. A kvalitatív adatok elemzését és kiértékelését követően dolgoztam ki azt a kérdőívet, illetve választottam ki azokat az itemeket, melyek a kvalitatív elemzés kiindulópontjai voltak. A kvantitatív kutatás feltárja az esetleges mérési hibákat, a kvalitatív módszerek célja pedig az eltérő itemműködés és az esetleges torzítások okainak feltárása, és olyan eljárásmódok kidolgozása, melyek a torzítás kiküszöbölésére alkalmasak, és ezáltal hozzájárulnak a mérés érvényességének növeléséhez.

Annak érdekében, hogy minél megbízhatóbb eredmény szülessen, a trianguláció módszerét is alkalmaztam (Flick, 2008; Sántha, 2013). A kérdőíves módszert kiegészítettem interjúkkal, melyekben lehetőségem volt, a nyelvtanárok és mérési szakértők mellett, a nyelvtanulók, vagyis a valós és a potenciális vizsgázók véleményének megismerésére is. A különböző forrásokból származó eredményeket egymást követően és összekapcsolva is értékeltem és elemeztem. A súlyozást tekintve a kvalitatív és a kvantitatív módszertan ugyanazzal a súllyal vesz részt a kutatásban.

6. A kutatás összefoglalása és fő eredményei

Az értekezés egyik célja az alábbi hipotézis tesztelése és az ehhez kapcsolódó kutatási kérdés megválaszolása volt:

> H: Az ECL B2-es szintű német nyelvi beszéd- és szövegértés feladatokban az éles vizsgát megelőző és azt követő minőségbiztosítási eljárások sokfélesége miatt csak korlátozott számú, a nemek szempontjából eltérő működést mutató item található, az itemek jelenléte azonban nem okoz szignifikáns különbséget a férfi és női vizsgázók feladatszintű teljesítményében.
>
> K1: Felfedezhető-e a beszéd- és szövegértés feladatokban eltérő itemműködés a vizsgázók neme szempontjából? Ha igen, milyen mértékben?

A hipotézis tesztelése és a kutatási kérdés megválaszolása a DIF-elemzés és a Wilcoxon-féle rangpróba kombinációjával valósult meg. A DIF-elemzés feltárta azokat az itemeket, amelyek esetén szignifikáns különbség mutatkozik a férfi és a női vizsgázói csoportok között, a Wilcoxon-féle rangpróba pedig azt vizsgálta, hogy az item szinten észlelt eltérő itemműködés hatással van-e a két vizsgázói csoport teljes teszten elért eredményére.

A DIF-elemzés a beszédértés itemek esetén összesen 13 eltérő itemműködést jelző itemet azonosított, a vizsgált időszakban az itemek 6,5%-a minősült DIF-nek. Hat item bizonyult könnyebbnek a férfiak számára, hét item pedig a nőknek kedvezett. Zwick és társai (1999) szerinti kategorizálás alapján a kutatás a férfiak esetén három markáns és három kevésbé markáns DIF, a nők szempontjából pedig öt markáns és két kevésbé markáns DIF található a feladatokban.

A vizsgált időszakban a szövegértés itemek 8,5 %-a (17 item) bizonyult DIF-nek. Az eredmények alapján a férfiak szempontjából kilenc item bizonyult szignifikánsan nehezebbnek, a nők esetén pedig nyolc item megválaszolása okozott nagyobb nehézséget. A DIF-mértékét tekintve mindössze egy item értéke minősül markáns DIF-nek, 16 item a kevésbé markáns DIF kategóriába sorolható.

A férfiak és a nők teszt szintű eredményeinek összehasonlítása Wilcoxon-féle rangpróba segítségével történt, a statisztikai eredmények nem jelzetek szignifikáns különbséget a férfiak és a nők teljesítménye között egyik vizsgaidőszakban sem. A DIF-elemzés és a Wilcoxon-féle rangpróba eredményei tehát igazolják a hipotézist, miszerint az ECL B2-es szintű német nyelvi beszéd- és szövegértés feladatokban valóban fellelhetőek korlátozott számú eltérő itemműködést

tanúsító itemek, ezek jelenléte azonban a teszt szintjén nem okoz szignifikáns különbséget a férfi és női vizsgázói csoportok sikerességében, és ezáltal nem befolyásolja a vizsga kimenetelét.

Az értekezés második célja olyan esetleges szisztematikus mintázatok felfedése volt, például az itemek nehézsége, az itemek általános illeszkedési mutatói, a feladattípusok és az itemek tartalma terén, amelyek szorosan kapcsolódnak az itemek eltérő működéséhez a nemek szempontjából, és potenciális forrásai lehetnek a torzításnak. E cél elérése érdekében két további kutatási kérdés fogalmazódott meg:

K2: Felfedezhetőek-e olyan szisztematikus mintázatok, amelyek rendszeresen az egyik csoportot hozzák előnybe, és az eredmények helytelen értelmezéshez vezethetnek?

K3: A DIF-ként jelölt itemek ténylegesen előnyben részesítik, vagy hátrányosan érintik a vizsgált csoportok egyikét, és ezáltal torzítják az eredményt?

A második kutatási kérdés megválaszolása további kvantitatív és kvalitatív kutatási módszerek bevonásával valósult meg. Az itemek nehézségének és diszkriminációs mutatóinak, valamint a teszt megbízhatóságának vizsgálata klasszikus tesztelméleti módszerrel történt. Az elemzés nem tárt fel egyik készség esetén sem szélsőségesen nehéz vagy könnyű itemeket. Beszédértés esetén két item bizonyult az elvártnál kissé könnyebbnek, szövegértés esetén egy item volt kissé nehéz, egy pedig kissé könnyű. Diszkrimináció szempontjából is jól működőnek bizonyultak az itemek, mindössze egy beszéd- és egy szövegértés item diszkriminációs mutatója marad el minimálisan az elvárt értéktől. A tesztek megbízhatósága (Cronbach-féle α értéke) minden esetben megfelelőnek bizonyult.

Az itemek általános illeszkedését a *Facets* program segítségével vizsgáltam. A szövegértés itemek illeszkedési mutatója minden esetben megfelelőnek bizonyult, mindössze két beszédértés item illeszkedési mutatója nem felelt meg az elvárásoknak, mindkettő megoldása a nőknek okozott nagyobb nehézséget. Az érintett itemeket azonban az éles vizsga után alkalmazott utólagos korrekciós eljárás kizárta az eredményszámításból. Mivel az említett két item semlegesítve lett, a beszédértés esetén a DIF-itemek aránya a vizsgált kétéves időszakban 5,5%-ra csökkent.

A második kutatási kérdés megválaszolása érdekében a beszédértés és szövegértés feladatok feladattípusát is vizsgáltam. A beszédértésben két feladattípus fordul elő: feleletválasztós és rövid válasz típusú feladat. A DIF-ként azonosított 13 item közül négy feleletválasztós, és a többi kilenc rövid válasz típusú volt. Az eredmények alapján nem lehet egyértelműen kijelenteni, hogy az egyik itemtípus rendszeresen előnyben helyezné valamelyik

nemet. Az itemek között kiegyensúlyozott eloszlás mutatkozott a férfiak és nők között, mind a feleletválasztós, mind a rövid válasz típusú feladatok esetében.

A szövegértés területén három típusú feladattal mérik a vizsgázók tudását, ezek a következőek: párosítás, mondatrészek visszahelyezése a szövegbe, valamint rövid válasz típusú feladatok. A feladatok típusát illetően kilenc item rövid válasz típusú, négy-négy item pedig párosítás és mondatrészek visszahelyezése a szövegbe típusba sorolható. A nemek közötti eloszlás alapvetően a szövegértés esetén is kiegyensúlyozott, azonban a rövid válasz típusú feladatok esetében észrevehető, hogy a női vizsgázók általában jobban teljesítettek.

Összességében megállapítható, hogy mind a beszéd-, mind a szövegértés feladatoknál tapasztalható, hogy a rövid válasz típusú feladatok esetén nagyobb valószínűséggel fordul elő DIF. A szövegértés esetén a vizsgált rövid válasz típusú itemek kedveznek a női vizsgázóknak, míg a többi feladattípus esetén nem mutatható ki egyértelmű tendencia a nemek közötti különbségek tekintetében.

Annak érdekében, hogy az itemek tartalmi szintjén megjelenő esetleges szisztematikus mintázatokról is részletes képet kapjak, az itemek tartalomelemzésére kérdőíves módszerrel és mélyinterjúk segítségével került sor. A kérdőív eredményeinek elemzése a következő felismerésekhez vezetett:

(1) Az ítészek véleménye és a statisztikai elemzések eredményei az esetek többségében összhangban vannak.

(2) Az ítészek többnyire úgy ítélik meg, hogy nincs olyan tartalom az itemekben, amelyik egyértelműen előnybe vagy hátrányba hozná az egyik csoportot, és torzítást okoz.

(3) Az ítészek azokban az esetekben, amikor úgy ítélik meg, hogy az item előnyben részesítheti az egyik nemet a másikkal szemben, alapvetően a témakörre és a szövegek tartalmára hivatkoznak. Ugyanakkor a statisztikai elemzések rámutatnak arra, hogy az itemek megoldása nem csupán a témától függ, és a vizsgázói csoportok nem feltétlenül úgy teljesítenek, ahogyan az ítészek feltételezik.

(4) Az ítészek indoklásaiból kiderül, hogy ritkán veszik figyelembe a nyelvi nehézségeket vagy kognitív folyamatokat a véleményük kifejtésekor. Leginkább a témát tekintik mérvadónak.

(5) Az ítészek úgy vélték, hogy például a gyermeknevelés, háztartás, híresség és emberi kapcsolatok témakörei előnyösebbek lehetnek a nők számára, míg a sport, technológiai és

vállalkozással kapcsolatos témakörök a férfiakat segíthetik. Azonban a statisztikai elemzés sok esetben nem erősítette meg ezt a véleményt. Ezen eredmények is azt támasztják alá, hogy előítéletek és sztereotípiák is befolyásolhatják az itemek besorolását.

(6) Az eredmények azt mutatják, hogy bizonyos témakörök esetén vannak olyan tényezők, amelyek befolyásolhatják a férfiak és nők megoldási képességét, ilyenek például az érdeklődési kör és a tapasztalat.

Az eredményeket összegezve azt látjuk, hogy az ítészek véleményének és a statisztikai elemzéseknek a kombinálása hatékony módszer lehet a DIF-itemek azonosítására, de nyilvánvalóvá válik az is, hogy például mélyinterjúk további információkkal szolgáltathatnak az eltérő itemműködés okairól és hozzájárulhatnak a DIF-itemek kiküszöbölésére szolgáló módszerek kidolgozásához.

A mélyinterjúk eredményei a következő összegzésekben foglalhatóak össze:

(1) Az interjúalanyok többsége az itemeket a nemek szempontjából semlegesnek ítéli meg, a megkérdezettek véleménye szerint a vizsgázó neme nem befolyásolja az item megoldását.

(2) Az interjúalanyok véleménye és a statisztikai eredmények közötti különbségek azt sugallhatják, hogy a valóságban a nemi különbségek hatása kevésbé mérvadó a válaszadásban, mint ahogyan azt az interjúalanyok egy része érzékeli.

(3) Az interjúalanyok általánosan úgy vélik, hogy a nemek közötti különbség nem játszik döntő szerepet a feladatok megoldásában. A szöveg összetettsége, a benne található specifikus kifejezések, a kérdések megfogalmazása, és leginkább a vizsgázók nyelvtudása, befolyásolhatja a vizsgázók teljesítményét.

(4) Az itemek megoldásának nehézsége nem a nemek közötti különbségekkel, hanem gyakran a szókincs ismeretével, a logikai gondolkodással és a témához való kapcsolódással van összefüggésben.

(5) Az interjúalanyok véleménye alapján vannak olyan szisztematikus mintázatok, amelyek felfedezhetőek az eltérő itemműködést tanúsító itemek tartalmában.

(6) Az interjúalanyok által felvetett javaslatok, például a kulcsszavak magyarázatának hozzáadása vagy a szövegek, kérdések módosítása, segíthetnek az itemek minőségének javításában.

Az eredmények szempontjából az ötöst pont kiemelkedő fontosságú, ugyanis az interjúalanyok véleménye szerint szisztematikus mintázatokat is fel lehet fedezni az itemek struktúráját illetően. A DIF-itemek tartalomelemzése az alábbi közös jellemzőket tárta fel:

(1) A válasz helyessége egyetlen lexikai elem megértésén múlik (lásd pl. beszédértés 5., 6., 9. item; szövegértés 17. item). Az esetek többségében határozószavakról van szó, amelyek figyelmen kívül hagyása helytelen válaszhoz vezet.

(2) A mérni kívánt szintet meghaladó lexikai elemek (pl. idiomatikus kifejezések és állandósult szókapcsolatok) és komplex struktúrák jelenléte, amelyek megértése elengedhetetlenül szükséges az item helyes megoldásához (lásd pl. beszédértés 2. és 3. item; szövegértés 2., 6., 7. és 8. item).

(3) Erős disztraktor jelenléte. A párosítás feladatok esetében ezek olyan válaszlehetőségek, amelyek látszólag helyesek, és csak amiatt zárhatóak ki, mert egy másik itemhez jobban illenek. Jellemzően ezek olyan tartalmi összefoglalók, amelyek nem a teljes szövegre vonatkoznak, hanem annak egy részére, ami megzavarhatja a vizsgázókat (lásd pl. szövegértés 14. és 16. item).

(4) Beszédértés esetén a válasz a hangzós szövegben előbb hangzik el, minthogy a vizsgázó a kérdésben és a szövegben található szemantikai átfedések segítségével észlelné, hogy a helyes válasz következik (lásd pl. beszédértés 1. és 8. item).

(5) A helyes válaszadáshoz minimum két információra van szükség (lásd pl. beszédértés 4. item; szövegértés 1. item).

(6) Nem egyértelmű szemantikai átfedés a kérdés és a szöveg között. Ezekben az esetekben a válaszadás logikai kapcsolat létrehozását feltételezi és a vizsgázó elbizonytalanodásához vezethet (lásd pl. beszédértés 7., 10. és 12. item.; szövegértés 9. item).

A második kutatási kérdés megválaszolása céljából végzett kvantitatív és kvalitatív elemzések eredmények alapján tehát megállapítható, hogy az itemek nehézsége és diszkriminációja nem mutat szélsőséges eltéréseket a nemek között, azaz általában nem tapasztalható szisztematikus előny vagy hátrány a nemek szempontjából. A vizsgált DIF-itemek általános illeszkedése megfelelő mind a beszédértés, mind a szövegértés terén. Bár két beszédértés item esetében észlelhető volt nem illeszkedés, ezeket az itemeket az utólagos korrekció segítségével az ECL Vizsgaközpont semlegesítette, így nem befolyásolják a vizsga kimenetelét és az eredmények megbízhatóságát. A feladattípusok tekintetében nem mutatható ki egyértelmű előny vagy hátrány egyik nem javára sem. Az itemek struktúrája terén azonban

tapasztalhatóak szisztematikus mintázatok. Az adatok elemzése során feltárt mintázatok segítenek megérteni, hogy milyen tényezők vannak hatással az itemek nehézségére. Ezek alapján megállapítható, hogy a lexikai elemek nehézsége, az erős disztraktorok jelenléte, a kérdés és a válasz közötti szemantikai kapcsolat, ill. a vizsgázók nyelvtudásának szintje nagyobb hatással van a helyes válaszadásra, mint a vizsgázók neme. Az eredmények alapján nincs empirikus bizonyíték arra, hogy a bármelyik itemtípus szisztematikusan előnyös vagy hátrányos lenne valamelyik vizsgázói csoport számára, és ezáltal veszélyeztetné a vizsga validitását és megbízhatóságát.

A harmadik kutatási kérdés megválaszolásához, szükség volt annak feltárására, hogy a DIF-ként jelölt itemek valóban eredménytorzítást okoznak-e a különböző nemű vizsgázók esetén:

K3: A DIF-ként jelölt itemek ténylegesen előnyben részesítik, vagy hátrányosan érintik a vizsgált csoportok egyikét, és ezáltal torzítják az eredményt?

A harmadik kutatási kérdést részben a második kérdés összegzése is megválaszolja. Annak érdekében azonban, hogy pontosabb képet kapjunk az itemek működéséről, kiválasztottam hét rövid válasz típusú beszéd- és öt szövegértés itemet, melyek esetében a kérdésekre adott vizsgázói válaszokat is elemeztem. Az eljárás célja az eltérő itemműködést okozó és lehetséges torzításhoz vezető tényezők feltárása volt, a két vizsgázói csoport helytelen válaszai elemzésének segítségével.

Az eredmények alapján az 1. és 3. beszédértés itemek esetén észlelhető torzítás. A tartalomelemzés ugyanis feltárta, hogy az itemek a szöveg ismerete nélkül is megválaszolhatóak, a kérdéseket a férfi vizsgázók pedig vélhetően a témával kapcsolatos háttértudásuk segítségével válaszolták meg. A többi vizsgált beszéd- és szövegértés item esetén az item nehézségét a lexikai hiányosságok, a szöveg értelmezésének nehézségei, a kulcsszavak figyelmen kívül hagyása, valamint a szöveg komplexitása okozta, és nem a nemek közötti különbségek. A helytelen válaszok eloszlásában nem észlelhető lényeges különbség a két vizsgázói csoport között.

7. Tanulságok és következtetések

A kutatás eredményei arra utalnak, hogy a vizsgát megelőző kvalitatív eljárások ellenére a nyelvvizsga feladatokban az empirikus vizsgálat olyan itemek jelenlétére utal, amelyek eltérő itemműködést tanúsítanak a vizsgázók nemek szerinti csoportjaiban. Ezeken az itemeken végzett további empirikus elemzéssel és kvalitatív módszerekkel nyert eredmények arra engednek következtetni, hogy ugyan a statisztikai elemzés a nemek szempontjából eltérő itemműködést jelez, a tartalomelemzés alapvetően nem erősíti meg a torzítás jelenlétét a feladatokban. Mindössze két esteben (beszédértés 1. és 3. item) utalnak arra a vizsgázói válaszok, hogy vélhetően a férfiak háttértudása az adott témában segítette az itemek megoldását. Ugyanakkor fontos megemlíteni, hogy az adott itemek esetén teszthibára is fény derül, az itemek ugyanis a témáról való ismeretek alapján, a szöveg meghallgatása nélkül is megválaszolhatóak, ez pedig veszélyezteti a validitást (Grotjahn, 2000: 47). Minden más esetben a kiegészítő elemzések arra utalnak, hogy az item nehézségét más tényező okozta. Vélhetően ez az oka annak is, hogy a nagymintás kérdőívet kitöltő ítészek többsége a szövegértés 11. item kivételével, minden itemet semlegesnek vélt a nemek szempontjából. A Likert-skála értékeinek átlaga tehát csupán egyfajta tendenciát mutat, hogy az ítészek egy kisebb csoportjának véleménye szerint, melyik nemet hozza előnybe vagy hátrányba az adott item. Amennyiben ezt a tendenciát összevetjük a DIF-elemzés eredményeivel megállapíthatjuk, hogy a vizsgált itemek 63%-ában (8 beszédértés és 11 szövegértés item esetén), az ítészek véleménye összhangban van a statisztikai elemzés eredményeivel. Az indoklások vizsgálata során azonban nyilvánvalóvá válik, hogy az ítészek leginkább a feladatok témája alapján hozták meg arra vonatkozó döntésüket, hogy az item melyik vizsgázói csoportnak kedvezhet. A döntéshozást tehát alapvetően sztereotípiák és előítéletek határozták meg, mintsem az item tartalmának, esetleg a szövegnek a vizsgálata.

A mélyinterjúk során az interjúalanyok az item és a szövegek tartalmára fokuszálnak, ezáltal pontosabb képet kapunk az itemek működéséről, az item nehézségét okozó tényezőkről és a stratégiákról. Az interjúk fényt derítenek az itemek közös jellemzőire is, amelyek alapján egyfajta kategorizálás is lehetséges.

A vizsgált itemek kapcsán tehát megállapítható, hogy a feladatok valóban tartalmaznak DIF-itemeket, melyek megoldásában szignifikáns különbség mutatkozik a nők és a férfiak csoportja között, de a tapasztalt különbségek nem magas arányúak, és nem befolyásolják érdemi módon az eredményt. Ezt bizonyítja az a tény is, hogy a vizsgált 30 itemből mindössze 8 item esetén beszélhetünk markáns DIF-ről. Ebből két item kapcsán (beszédértés 1. és 3. item)

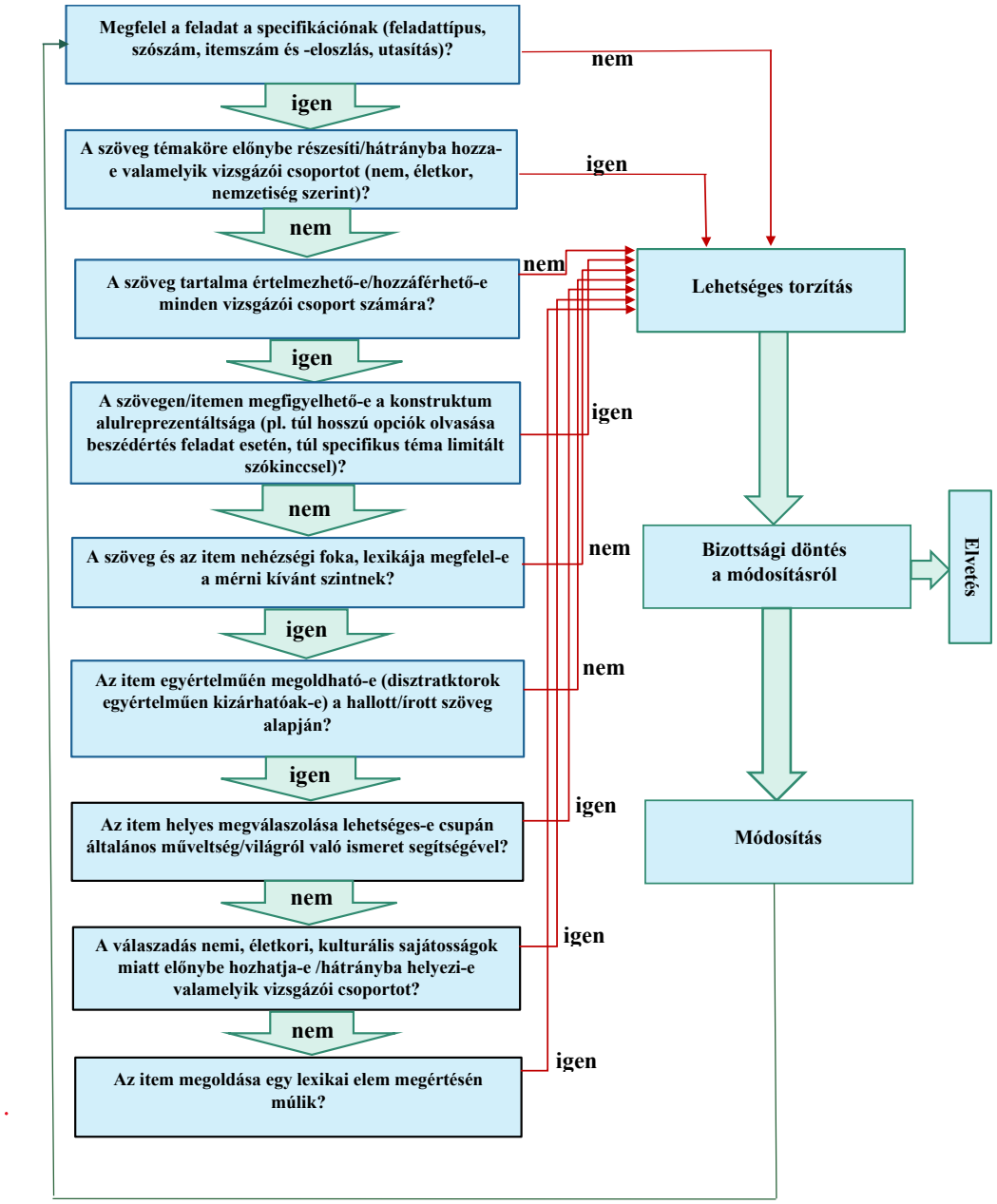
beazonosítható a DIF forrása, ami az eredmények torzításához vezethetett. A többi item esetén az eltérő itemműködés okára ugyan egyértelműen nem derült fény, de a kutatás során kizárhatóak voltak olyan tényezők, mint a feladattípus, az itemek általános nehézsége és diszkriminációs értéke, a feladatok megbízhatósága, amelyek az értekezés eredményei alapján nem okoznak mérvadó különbséget a két vizsgázói csoport teljesítményében.

Az itemek megoldását leginkább nyelvi hiányosságok gátolhatják, ezek általában szókincsbéli hiányosságok, amelyek vélhetően nem nem- vagy témafüggőek, több esetben határozószavakról vagy egyszerű kérdőszavakról van szó. Mivel nem valószínű, hogy ezek a hiányosságok kizárólag egy adott vizsgázói csoportra jellemzőek, ezeket nem tekinthetjük egyértelműen az eltérő itemműködés okának. Említésre méltó ugyanakkor az a tény, hogy ugyan az elemzések a teljes vizsgázói populáció eredményeit veszik figyelembe, a populáció összetétele véletlenszerű, és sok esetben nem magas létszámú (nem reprezentatív), így valószínűtlen válaszmintázatok is előfordulhatnak, amelyek eltérő itemműködésben nyilvánulhatnak meg a két csoport viszonylatában. Ennek felderítése további kutatások tárgyát képezheti.

Az értekezés célja nem mindössze az eltérő itemműködést tanúsító itemek és azok esetleges közös tulajdonságainak feltárása volt, hanem a különböző vizsgát megelőző és az éles vizsgát követő kvantitatív és kvalitatív módszerek hatékonyságának vizsgálata is. Az eredmények alapján nyilvánvalóvá válik, hogy a tesztfejlesztés folyamán nagyobb figyelmet kell fordítani az itemek kidolgozására. Az itemek írásakor és a bizottsági munka során is az eltérő itemműködést kiküszöbölésére irányulóan konkrét ajánlásokra van szükség a nyelvvizsga fejlesztői számára.

A kutatás eredményeként és a fenti cél megavalósításaként egy modellt (lásd 1. ábra) hoztam létre, amely a tesztírás szakaszában alkalmazható. A modell az eltérő itemműködésre fókuszál, de nemcsak a nemek, hanem az életkor, nemzetiség és kulturális háttér szempontjából is vizsgálja a jelenséget. Fejlesztése során figyelemebe vettem az ALTE által kiadott tesztírói kézikönyvben megfogalmazott feladat- és itemírásra vonatkozó általános szempontokat (ALTE, 2011: 66), illetve az Allalouf és társai (1999: 196) által vázolt DIF megállapítására kifejlesztett folyamatábrát, melynek fókuszában a fordítások során kialakuló eltérő itemműködés elkerülése áll.

1. ábra Az eltérő itemműködés megállapításának modellje



A kidolgozott modell a szöveg szintjétől halad lefelé az item megfogalmazásáig, figyelembe veszi a kutatás során feltárt problémákat, és a beszéd- és szövegértés feladatokra fókuszál. Néhány apróbb módosítás után beszéd- és íráskészség feladatok fejlesztése során is alkalmazható, hiszen a produktív készségek esetén a nemből, korból, kulturális háttérből adódó előnyök és hátrányok jelentősen befolyásolhatják a feladatok megoldását, és ezáltal veszélyeztetik a méltányosság elvét.

Annak érdekében, hogy a tesztírók, tesztfejlesztők és az ítészek hozzájáruljanak ahhoz, hogy minőségi tesztek készüljenek, nem csupán a tesztfejlesztés szakaszában, hanem a kipróbálás és az éles vizsgán való felhasználás után is szükséges van folyamatos visszacsatolásra

a feladatok működéséről. Az ECL német vizsga esetében is különösen fontos ez, hiszen az elmúlt két évben a vizsgázói populáció lényegesen megváltozott, ami a tesztírókat is új kihívások elé állítja, az itemírást ugyanis jelentősen megnehezíti, ha a vizsgázók heterogének (Grotjahn, 2000: 47). A nagyobb kihívást tehát vélhetően nem a fent említett szempontok figyelembevétele okozza majd, hanem egy olyan nemzetközi ítészi bizottság létrehozása, amelyik kellő ismerettel rendelkezik a tesztírásról és a vizsgázók kulturális hátteréről. A receptív készségek (beszéd- és szövegértés) szempontjából ennek kisebb a relevanciája, mivel a vizsgázók az itemeket az elhangzó vagy leírt szövegek alapján oldják meg, a szövegek minden olyan információt tartalmaznak, amelyekre a helyes válasz megtalálásához szükség van. A produktív készségek (beszéd- és íráskészség) esetén, ahol a vizsgázónak önállóan kell megfogalmazniuk véleményüket egy adott témával kapcsolatosan, komoly hátrányt jelent, ha nem rendelkeznek valamennyi tudással vagy tapasztalattal az adott témáról. Ebben az esetben a hozzáférhetőség elve is sérül (AERA et al., 2014), ami szintén okozhat eltérő itemműködést, és ezáltal veszélyezteti a validitást. Ennek kiküszöbölése érdekében, az eltérő itemműködés vizsgálatára lehetőség szerint már az empirikus validálási szakaszban is szükség van.

A megnövekedett vizsgázói létszám magasabb preteszt létszámokhoz is vezet, így lehetőség nyílik az eltérő itemműködés vizsgálatára már az empirikus validálás szakaszában is. Így a fent említett visszacsatolás már az éles vizsga előtt is megtörténhet. Ha az éles vizsgát megelőzően is lehetőség nyílik az eltérő itemműködés vizsgálatára, valamint a tesztírás, a tesztfejlesztés, az értékelés és bizottsági ülések során szerzett tudás és tapasztalatok is beépültnek a tesztelési folyamatba, vitathatatlanul javul a feladatok minősége, amely pozitív hatással lesz a tesztek megbízhatóságára és a vizsga validitására. Az értekezésben feltárt mintázatok figyelembevétele és ezek elkerülése a tesztírás során, elősegíti a méltányosság megvalósulását. A modell segítheti a tesztfejlesztésben résztvevő kollégák munkáját, és tudatosítja, hogy a méltányosság egy magas kockázatú vizsga esetén kiemelt figyelmet érdemel. A fő cél ugyanis a szisztematikus DIF kizárása (Du, 1995) és annak biztosítása, hogy egyenlő eséllyel induljanak a nyelvvizsgán a különböző nemű, korú és kulturális háttérrel rendelkező vizsgázók.

8. Kitekintés

A DIF-elemzésnek kulcsfontosságú szerepe lesz az ECL Vizsgaközpont jövőbeli kutatásaiban is. A nemzetközi szinten német nyelvből B1 és B2 szinten jelentősen megnövekedett vizsgázói létszám szükségessé teszi a DIF-elemzés kiterjesztését és az eltérő itemműködés vizsgálatát a különböző nemű, korú és anyanyelvű vizsgázói csoportokban. Az értekezés fókuszában a receptív készségek (beszéd- és szövegértés) állnak, azonban kiemelten fontos lenne a produktív készségek (beszéd- és íráskészség) feladatainak vizsgálata is. A kidolgozott kutatási modell más készségekre is adaptálható és több változóra is kibővíthető.

Az eredmények alapján számos kutatási irány merül fel a beszéd- és szövegértés itemek továbbfejlesztése és javítása érdekében. A DIF-elemzés segítségével vizsgálható például az olyan tényezők hatása is az eredményre, mint az oktatási vagy a kulturális háttér. Továbbá a kutatás kiterjeszthető más vizsganyelvekre is. Ezáltal vizsgálható lenne, hogy azok a mintázatok, amelyek a német nyelvi tesztek mutattak, nyelvspecifikusak-e vagy általánosíthatóak-e az eredmények. A több nyelvre kiterjedő kutatás átfogóbb képet adna a jelenségről, és vélhetően új eltérő itemműködést és torzítást megelőző módszerek kidolgozását segítené elő.

Jelen értekezésben az ítészek által kitöltött kérdőív – a vizsgált itemek magas számára való tekintettel – kizárólag DIF-itemeket tartalmazott. Egy következő kutatás lehetővé tenné a DIF-itemek és a statisztikai eredmények alapján megfelelően működő itemek együttes elemzését, és ezáltal az ítészek megbízhatóságának és a statisztikával való összhangjuknak a vizsgálatát.

Fontos lenne annak vizsgálata is, hogy az autentikus, a valós életben előforduló témák és tartalmak milyen mértékben módosíthatóak az eltérő itemműködés elkerülése céljából, anélkül, hogy a módosítások veszélyeztetnék a tartalmi érvényességet.

A kutatás eredménye sok esetben arra utal, hogy lexikai hiányosságok, logikai kapcsolatok létrehozásnak, illetve a közös szemantikai tartalmak beazonosításának nehézségei is okozhatnák az eltérő itemműködést. Megfontolandó tehát ennek a vizsgálata és a nyelvkönyvek tartalmának elemzése olyan szempontból, hogy ezeket a képességeket kellő módon fejlesztik-e a használatban lévő tankönyvek. Az eredmények alapján olyan nyelvtanulásra és vizsgafelkészítésre is alkalmas tananyagrészek fejlesztésére nyílna lehetőség, amelyek segítenek a diákoknak bővíteni a szókincsüket és fejleszteni nyelvi képességeiét.

A nyelvvizsgákon alkalmazott itemek és tesztek érvényességének, megbízhatóságának és méltányosságának fenntartása folyamatos feladat, amely az újabb kutatások és ezek eredményének figyelembevételével valósulhat meg.

Irodalomjegyzék

2012. évi 423. (XII. 29.) Korm. rendelet a felsőoktatási felvételi eljárásról, 19.§ Elérhető: https://net.jogtar.hu/jogszabaly?docid=a1200423.kor Letöltés dátuma: 2023.július 17.

AERA, APA & NCME. (1999). *Standards for Educational and Psychological Testing.* Washington DC: American Educational Research Association Publishing.

AERA, APA & NCME. (2014). *Standards for Educational and Psychological Testing.* Washington DC: American Educational Research Association.

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

Association of Language Testers in Europe (ALTE). (1998). *Multilingual Glossary of Language Testing Terms.* Cambridge: Cambridge University Press.

Association of Language Testers in Europe (ALTE). (2011). *Manual for Language Test. Development and Examining. For use with the CEFR.* Strasbourg: Council of Europe.

Association of Language Testers in Europe (ALTE). (2020). *ALTE Principles of Good Practice.* Cambridge: ALTE.

Auswärtiges Amt. (2023). Visumhandbuch. Letöltés dátuma: 2023. július 22., forrás: https://www.auswaertiges-amt.de/blob/207816/246dbf01c23d5e51ba5790ed0d4c811a/visumhandbuch-data.pdf

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing.* Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language Testing in Practice. Designing and Developing Useful Language Tests.* Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world.* Oxford, UK: Oxford University Press.

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

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items.* Thousand Oaks, CA: SAGE Publications.

Chen, Z., & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing*, 155–163.

Cole, N. S. (1997). *The ETS gender study: How females and males perform in educational settings.* Princeton, NJ: Educational Testing Service.

Coniam, D., & Lee, T. (2021). *Potential bias in LanguageCert IESOL items: A Differential Item Functioning analysis.* London, UK: LanguageCert.

Creswell, J. W. (2004). *Research design: Qualitative, quantitative, and mixed methods approaches* (második. kiad.). Thousand Oaks, CA: Sage Publications.

Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and Conducting Mixed Methods Research.* Thousand Oaks: Sage Publications Ltd.

Cronbach, L. J. (1984). Essentials of psychological testing. New York: Harper & Row Publishers.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52(4)*, 281–302.

Dörnyei, Z. (2007). *Research methods in applied linguistics.* Oxford: Oxford University Press.

Eckes, T. (2011). A study of differencial item functioning in the TestDaF Reading and Listening section. In E. D. Galaczi (Szerk.), *Exploring Language Frameworks: Proceedings of the Alte Kraków Conference, July 2011.*

Elder, C. (2012). Bias in Language Assessment. Letöltés dátuma: 2020. szeptember 25, forrás: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781405198431.wbeal1198

Educational Testing Service. (ETS). (2016). ETS International Principles for the Fairness of Assessments. A Manual for Developing Locally Appropriate Fairness Guidelines for Various Countries. Letöltés dátuma: 2021. január 28, forrás: https://www.ets.org/s/about/pdf/fairness_review_international.pdf

Fekete, A., & Csépes, I. (2018). B2-es szintű nyelvvizsga bizonyítvány: útlevél a diplomás élethez, társadalmi mobilitáshoz. *Iskolakultúra, 28. évfolyam, 10-11. szám*, 13-24.

Fényes, H. (2008). Kontextuális hatások a középiskolások eredményességére. *Szociológiai Szemle, 18(3)*, 3-31.

Fényes, H. (2009). Nemek szerinti iskolai eredményesség és a férfihátrány hipotézis. *Magyar Pedagógia, 109 (1)*, 77-101.

Fényes, H. (2010a). *A nemi sajátosságok különbségének vizsgálata az oktatásban. A nők hátrányainak felszámolódása?* Debrecen: Debrecen: Debreceni Egyetemi Kiadó.

Fényes, H. (2010b). Horizontal and Vertical Segregation in Education by Gender in the Hungarian - Romanian - Ukrainian Border Region. *Journal of Social Research and Policy, 1(1)*, 49-68.

Ferne, T., & Rupp, A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly, 4*, 113-148.

Flick, U. (2008). *Triangulation. Eine Einführung.* Wiesbaden: VS Verlag.

Geranpayeh, A., & Kunnan, A. J. (2007). Differential Item Functioning in Terms of Age in the Certificate in Advanced English Examination. *Language Assessment Quarterly, 4 (2)*, 190-222.

Grotjahn, R. (2000). Studieneinheit Leistungsmessung und Leistungsbeurteilung. Patras: Hellenic Open University.

Hrisztova-Gotthardt, H., & Werner, R. (2021). Improving the fairness of ECL listening tests by detecting gender-biased items. *Strani jecizi. Vol. 50 (2)*, 207-234.

Hyde, J., & Linn, M. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin, 104(1)*, 53–69.

Joint Committee on Testing Practices. (2005). Code of fair testing practices in education (revised). *Educational Measurement: Issues and Practice, 24*, 23-29.

Kane, M. (2010). Validity and fairness. *Language Testing 27*, 177-182.

Kane, M. (2011). Validating score interpretations and uses: Messick Lecture, Language Testing Research Colloquium, Cambridge, April 2010. *Language Testing 29(1)*, 3–17.

Kane, M. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement, Vol. 50, No. 1*, 1–73.

Kispál, S., & Gergely, B. (2020). Eltérő itemműködés vizsgálata az országos kompetenciamérésben halmozottan hátrányos helyzetű diákok körében többdimenziós írt modellek segítségével. *Psychologia Hungarica Caroliensis, 8 (4)*, 150-179.

Kiszely, Z. (2022). Államilag elismert nyelvvizsgák tudományos feldolgozottsága magyarországi alkalmazott nyelvészeti és pedagógiai szakfolyóiratokban. *Iskolakultúra 32. évfolyam, 12. szám*, 22-40.

Kremmel, B. (2019). *Avoiding bias in language test development. (előadás az "ATLE 54th Meeting and Conference" rendezvényen).* Ljubljana.

Kunnan, A. J. (2000). Fairness and Justice for All. In A. J. Kunnan (Szerk.), *Fairness and validation in language assessment*. Cambridge: Cambridge University Press. 1-13.

Kunnan, A. J. (2004). Test Fairness. In M. Milanovic , & C. J. Weir (szerk.), *European language testing in a global context*. Cambridge: Cambridge University Press. 27-48.

Kunnan, A. J. (2007). Test Fairness, Test Bias, and DIF. *Language Assessment Quarterly 4 (2)*, 109-112.

Kunnan, A. J. (2014). Fairness and Justice in Language Assessment. In A. J. Kunnan (Szerk.), *The Companion to Language Assessment*. Hoboken, New Jersey: John Wiley & Sons. 1-17.

Mackey, A., & Gass, S. M. (2005). *Second language research: Methodology and design.* Mahwah, New Jersey: Lawrence Erlabaum Associates Publishers.

McNamara, T. (2000). *Language Testing.* Hong Kong: Oxford University Press.

McNamara, T., & Roever, C. (2006). *Language Testing: The Social Dimension.* Malden: Blackwell.

Messick, S. (1989). Validity. In R. L. Linn (Szerk.), *Educational measurement.* Washington, DC: American Council on Education and National Council on Measurement in Education. 13–103.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing*, *19*(4), 477–496.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus Article: On the Structure of Educational Assessments, Measurement. *Interdisciplinary Research and Perspectives*, 1:1, 3-62.

Pae, T. (2004). DIF for examinees with different academic backgrounds. *Language Testing, 21(1)*, 53-73.

Ross, S., & Okabe, J. (2006). The subjective and objective interface of bias detection on language testing. *International Journal of Testing, 6(3)*, 229-253.

Ryan, K. E., & Bachman, L. F. (1992). Differential item functioning on two tests of EFL proficiency. *Language Testing 9*, 12-29.

Sántha, K. (2013). A harmadik paradigma a neveléstudományi vizsgálatokban. *Iskolakultúra (2)*, 82-91.

Shepard, L. A. (1982). Definition of bias. In R. A. Berk (Szerk.), *Handbook of methods for detecting bias*. Baltimore: John Hopkins University. 9-30.

Song, X. (2014). *Differential Item Functioning investigations with Pearson English Test items.* London: Pearson.

Stoynoff, S. (2012). Fairness in Language Assessment. In C. A. Chapelle (Szerk.), *The Encyclopedia of Applied Linguistics (Online Edition).* Letöltés dátuma: 2020. szeptember 25, forrás:

https://onlinelibrary.wiley.com/doi/abs/10.1002/9781405198431.wbeal0409

Urbina, S. (2004). *Essentials of Psychological Testing.* Hoboken, New Jersey: John Wiley & Sons.

Vígh, T. (2005). A kommunikatív tesztelés elméleti alapjai. *Magyar Pedagógia, 105. 4. sz.*, 381-407.

Wesolowski, B., & Wind, S. (2019). Validity, Reliability, and Fairness in Music Testing. In T. S. Brophy (Szerk.), *he Oxford Handbook of Assessment Policy and Practice in Music Education. Vol 1*. Oxford: Oxford University Press. 436-460.

Zwick, R., Brown, T., & Sklar, J. C. (2004). California and the SAT: A Reanalysis of University of California Admissions Data. *Research and Occasional Papers Series*, 1-35.

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An Empirical Bayes Approach to Mantel-Haenszel DIF Analysis. *Journal of Educational Measurement, 36, 1*, 1-28.

Az értekezés témakörében releváns saját publikációk

Hrisztova-Gotthardt, H., Ugor, B., & Werner, R. (2016). *Thematisches Übungsbuch zur ECL Prüfungsvorbereitung Deutsch Stufe B2 Band 3*. Nyíregyháza: SZABÓ Nyelviskola és Fordítóiroda Kft.: SZABÓ Nyelviskola és Fordítóiroda Kft.

Hrisztova-Gotthardt, H., & Werner, R. (2017). Können Misfits vermieden werden? Eine quantitative und qualitative Analyse von nicht integrierbaren Items in Aufgaben zum Hörverstehen. In к. и.-В. Департамент по чуждоезиково обучение (Szerk.), *Чуждият език и съвременното висше образование. Сборник с доклади от VIII международна конференция, 23–25. VI. 2017, Варна, България*. Варна: Медицински университет "Д-р П. Стоянов". 634-640.

Hrisztova-Gotthardt, H., & Werner, R. (2021a). Improving the fairness of ECL listening tests by detecting gender-biased items. *Strani jecizi. Vol. 50 (2)*, 207-234.

Hrisztova-Gotthardt, H., & Werner, R. (2021b). Noch mehr Chancengleichheit für Prüfungsteilnehmende mit spezifischem Bedarf: Ein gemeinsames Vorhaben der Ungarischen Akkreditierungsbehörde für Sprachprüfungen und des ECL Sprachprüfungssystems. In ALTE (Szerk.), *Collated Papers for the ALTE 7th International Conference, Madrid*. Cambridge: ALTE. 124-129.

Werner, R. (2010). *Übungsbuch zur ECL Prüfungsvorbereitung. Deutsch Stufe B1*. Nyíregyháza: SZABÓ Nyelviskola és Fordítóiroda Kft.

Werner, R. (2018). Reliabilitätsanalyse der Bewerter der schriftlichen Kommunikation im Rahmen der ECL Sprachprüfung. In E. Zelenická (Szerk.), *Moderné jazyky v súčasnej Európe*. Nitra: Univerzita Konstantina Filozofa v Nitre. 148-156.

Werner, R. (2019). Megvalósul az esélyegyenlőség az ECL nyelvvizsgán? Diszgráfiás, diszlexiás és tanulási nehézségekkel küzdő vizsgázók teljesítményének elemzése. In J. T. Karlovitz, & J. Torgyik (szerk.), *Szakmódszertani és más emberközpontú tanulmányok.* Komárno: International Research Institute. 239-253.

Werner, R. (2019). Német nyelvű olvasásértés feladatok megbízhatóságának vizsgálata. In J. T. Karlovitz, *Tanulmányok a tanügy és az oktatástan világából*. Budapest: Neveléstudományi Egyesület. 87-97.