KISZELY ZOLTÁN

Értékelési skálák használata a beszédkészség és íráskészség vizsgákon

Rating scales in speaking and writing performance assessment in high-stakes examinations
While assessing examinees' speaking and writing performances in high-stakes examinations raters / examiners use rating scales. Rater effects like the halo effect, the severity effect, central tendency or inconsistency as construct-irrelevant factors should be minimized by language examination providers; therefore, it is of utmost importance to investigate similarities and differences in raters' interpretation of the rating scale. For this reason, the present study aims to compare the way in which the same group of examiners use the rating scales for the assessment of candidates' English B2 level performance in speaking and writing. The data were collected from the online retraining sessions organised for examiners at the BME Language Examination Centre in 2017 and 2018. First, the paper describes the most salient features of the rating scales for assessing speaking and writing performance. Second, it examines how stringently raters apply the different criteria of the two rating scales. Third, it discusses the similarities and differences in the use of the rating scales for evaluating candidates' production skills.

Keywords: productive skills, performance assessment, rating scales, rater effects, classical and modern test theory

Bevezetés

Minden Magyarországon akkreditált nyelvvizsga a vizsgázók négy nyelvi készségét méri és értékeli. Ezek közül a produktív készségek, vagyis a beszédkészség és az íráskészség értékelése olyan eljárás, amikor a vizsgázók kompetenciáját performanciájukon, azaz nyelvi teljesítményeiken keresztül vizsgálják (McNamara 1996). Az értékeléshez mérőeszközökre van szükség, azaz egy vagy több olyan feladatra, amelyek célja bizonyíthatóan az adott készség mérése, továbbá egy többnyire analitikus, azaz különböző értékelési szempontokat tartalmazó értékelési skálára, amelynek segítségével érvényes és megbízható módon lehet értékelni a teljesítményeket, és végül, de nem utolsósorban jól kiképzett értékelőkre, akik megfelelően használják az értékelési skálákat. A beszédkészség és az íráskészség ugyan eltérő konstruktumot takar, azaz nem ugyanazt méri, de mindkettő produktív készség, és az értékelés során használt eszközök és módszerek hasonlóak.

Felületesen gondolkodva lehetne azt állítani, hogy a skálahasználat egyértelműen annak a függvénye, hogy milyen nyelvi teljesítményt nyújtott a vizsgázó. Ha jól szerepelt, magas pontszámokat kap, ha gyengébben, akkor alacsonyabbakat. Ennek ideális esetben így is kellene lennie, de mivel humán értékelőkről van szó, nem csupán a vizsgázói nyelvi teljesítmény az egyetlen értékelést befolyásoló tényező, arra számos szubjektív elem is hat. Eckes (2009) jó néhány ilyen tényezőt sorol fel. Idetartozik például az értékelők szakmai háttere, a személyiségük és a leterheltségük, hogy csak néhányat említsünk. E szubjektív elemekkel terhelt értékelői hatásoknak (*rater effects*) a kiküszöbölése, illetve minimalizálása minden vizsgarendszer alapvető feladata, ugyanis az értékelői hatásoknak konstruktum-irreleváns (a mérés fókuszához nem tartozó) tényezőkként közvetlen következményük van az értékelés pontosságára vonatkozóan. A legfontosabb kérdés a skálahasználattal kapcsolatban tehát az, hogy a skála kritériumait megfelelő módon értelmezik-e a vizsgáztatók (Eckes 2019).

A fentiek miatt belátható, hogy az értékelés, azaz az értékelési skála használata olyan sokismeretlenes egyenlet, amelyet érdemes behatóbban is megvizsgálni. Mind a beszédkészség, mind pedig az íráskészség tekintetében a szakirodalom már részletesen foglalkozott az értékelési skálák és a vizsgáztatók interakciójával (Kuiken–Vedder 2014, Wind–Peterson 2017), de általában ezek a kutatások a két készséget külön-külön, nem egymással összehasonlítva vizsgálták. Jelen tanulmány ezért arra tesz kísérletet, hogy a BME Nyelvvizsgaközpont 2017. és 2018. évi vizsgáztató-továbbképzése során összegyűlt adatok segítségével összehasonlítsa, hogyan használják ugyanazon angol nyelvi vizsgáztatók az íráskészség és a beszédkészség méréséhez készült értékelési skálákat két-két B2 szintű, de különböző nyelvi színvonalú nyelvvizsgázói teljesítmény értékelése alapján. A tanulmány az értékelők, az értékelési skálák és a vizsgázók közötti interakcióval foglalkozik, de e három tényező közül elsősorban az értékelési skála használatára fókuszál, ebből a szempontból vizsgálja a kérdést. Annak ellenére, hogy az adatok vizsgáztató-továbbképzésekből származnak, a tanulmánynak nem célja annak bemutatása és elemzése, hogy a vizsgáztatók értékelése mennyire közelített a vizsgaközpont által felállított standardokhoz.

Az első részben rövid szakirodalmi áttekintést adunk az értékelési skálák használatával kapcsolatos korábbi kutatásokból. Ezt követően a kutatási kérdések és a választott módszertan leírása következik. Az elemzés során először a beszédkészség és az íráskészség értékelésére alkalmazott skálák használatának modern tesztelméleti eszközökkel kimutatott jellemzőit vizsgáljuk, majd a skálák egyes kritériumaival kapcsolatos vizsgáztatói szigort helyezzük a középpontba. Ezt követi annak vizsgálata, hogy milyen összefüggések és különbségek mutathatók ki a két skála használatában. A tanulmányban az *értékelő* és *vizsgáztató* szavakat, valamint az *értékelési kritérium* és az *értékelési szempont* kifejezéseket szinonimaként fogjuk használni.

Az eredményekből egyrészt hasznos tanulságok vonhatóak le a vizsgaközpont számára, illetve az eddigi esetleges megérzések konkrét adatokkal támaszthatók alá, másrészt pedig kirajzolódhat egy kép a mai nyelvtanári/nyelvvizsgáztatói szakma egy részének produktív készségekkel kapcsolatos látásmódját és attitűdjét illetően.

Szakirodalmi áttekintés

Ahogy a bevezetőben említettük, jelen tanulmány az értékelési szempontok felől közelíti meg az értékelés problematikáját, azonban a kritériumok, azaz a skála használatára a vizsgáztatói értékelésekből lehet következtetéseket levonni, ezért a szakirodalmi áttekintés során az értékelőket és a skála használatát nem lehet szétválasztani.

Az értékelői hatásokkal sokan foglalkoztak, de talán a legátfogóbb áttekintést Myford–Wolfe (2003) tanulmánya adja. Az általuk említett hatások közül a legfontosabbak a következők. Holdudvarhatásról (*halo effect*) akkor beszélünk, ha az analitikus értékelési skála egyik kritériumára adott pontszám túlságosan befolyásolja a többi kritériumra adott pontszámot, például, ha a nyelvhelyességi szempont gyenge pontszámot kap, akkor a szókincs is és a tartalom is automatikusan kevés pontot kap. Az elfogultság az egyik kritériummal vagy feladattal szemben azt jelenti, hogy túlságosan szigorú vagy engedékeny a pontozás az értékelés egy aspektusával kapcsolatban. A középső pontszámra törekvés tendencia (*central tendency*) azt jelenti, hogy a kockázatok elkerülése végett a vizsgáztatók nem használják a skála szélső értékeit, ezért értékeléseikben többnyire a középső pontszámok szerepelnek, például egy ötfokozatú (1–5) skála esetén a 3 a középső pontszám. A következetlenség az a jelenség, amikor egy-egy kritériumra egészen más pontszámot ad egy értékelő, mint a többi. E négy értékelői hatás közül kettőre, a holdudvarhatásra és a középső pontszámra törekvés tendenciára az értékelési szempontok felől is lehet következtetni, ahogy azt később látni fogjuk.

Az értékelési skálák validálásának statisztikai eszközeiről Eckes (pl. 2009, 2015, 2019) több tanulmányában is világos összegzést nyújt. A kritériumok modern tesztelméleti eszközökkel kimutatható adataiból egyrészt megállapítható, hogy a skála szerkezete a pontozás szempontjából megfelelően működik-e, másrészt az is kimutatható, hogy az analitikus skálák egyes kritériumai milyen mértékben járulnak hozzá az adott készség méréséhez, azaz az adott készséget mérik-e, vagy esetleg valami mást.

Az egyes kritériumokkal szembeni szigorúság kérdésével szintén több tanulmány foglalkozott. McNamara (1990) egy ausztrál egészségügyi dolgozók számára kifejlesztett szaknyelvi teszt validálásakor két különböző időpontban is azt találta, hogy mind a beszéd-, mind az íráskészség teljesítmények értékelésekor a nyelvtan a domináns, azaz a legszigorúbban értékelt kritérium. A TestDaF, német mint idegen nyelvi vizsga során megírt íráskészség teljesítmények elemzésekor Eckes (2009) kimutatta, hogy a nyelvhelyesség (*linguistic realisation*) volt a legszigorúbban értékelt kritérium.

Corrigan (2007) olasz beszédkészségvizsgán nyújtott nyelvi teljesítmények értékelése során szintén azt állapította meg, hogy a nyelvtani pontosság (*accuracy*) az a kritérium, amit az ítészek a legszigorúbban értékeltek, az interakciót és a folyékonyságot pedig a legenyhébben. Velük ellentétben Benke (2008) íráskészség-értékelők és -értékelési skála közötti interakciókat kutatva nem talált olyan kritériumot, amelyet konzekvensen szigorúbban vagy engedékenyebben értékeltek volna a vizsgáztatók. Ez arra utal, hogy az értékelők szigora vagy engedékenysége nagyon feladatfüggő, és azt is megerősíti, hogy szituációfüggő az értékelők értékelőskála értelmezése (Lumley–McNamara 1995).

Tudomásunk szerint a szakirodalom nem foglalkozott közvetlenül azzal a kérdéssel, hogy ugyanaz az értékelői csoport a két produktív készség nyelvi teljesítményeit menynyire hasonlóan vagy különbözően értékeli, azaz az értékelési skálákat hogyan használja, ezért jelen tanulmány erre a kérdésre fókuszál.

Kutatási kérdések

A vizsgálat három kérdésre kereste a választ, amelyek a következők voltak:

1. Mlyen jellemzőkkel rendelkeznek a BME B2 szintű beszédkészség- és íráskészség-értékelési skálái a vizsgáztatói értékelések tükrében?
2. Feladatonként a beszédkészség- és az íráskészség-teljesítményeket mely kritériumok mentén értékelték a legszigorúbban és a legenyhébben a vizsgáztatók 2017-ben és 2018-ban?
3. Milyen összefüggések és különbségek mutathatók ki a beszédkészség és az íráskészség értékelésére használt skálák alkalmazása között?

Kutatási módszerek

Résztvevők

A vizsgálat körébe 67, a BME általános nyelvvizsgarendszeréhez akkreditált angol nyelvi vizsgáztató értékeléseit vontuk be, akik 2017-ben és 2018-ban részt vettek mind beszédkészség, mind pedig íráskészség vizsgarészekhez szervezett online vizsgáztatótovábbképzéseken.

A továbbképzés során értékelt szóbeli és írásbeli nyelvi teljesítményeket minden évben az angol nyelvi felelősök választották ki. A kiválasztáskor a beszédkészség esetén a felvétel minősége mellett fontos szerepet játszott az is, hogy a két különböző évben különböző színvonalú teljesítmény legyen az értékelés tárgya, azaz ne hasonló pontszámú teljesítmények kerüljenek értékelésre.

Vizsgáztató-továbbképzés

A résztvevők 2017-ben és 2018-ban beszédkészség- és íráskészség-fókuszú online formátumú vizsgáztató-továbbképzésen vettek részt, amelynek során mindkét évben egy-egy vizsgázó beszédkészség- és íráskészség-teljesítményét értékelték a BME rendszerszerűen alkalmazott analitikus értékelési skálái segítségével.

A beszédkészség feladatlapon mindkét évben egy-egy B2 szintű vizsga hangfelvétele szerepelt, ami annyiban különbözött az éles vizsgától, hogy mivel audio anyagról volt szó, sem a vizsgázót, sem a vizsgáztatókat nem lehetett látni. A feladatlapon a vizsgán (azaz a hangzó anyagban) felhasznált feladatokat és az értékelési szempontokat is feltüntettük, így a hallgatás közben a feladatokat és a kritériumokat is egyszerre lehetett tanulmányozni. Az íráskészség feladatlapon mindkét évben egy-egy vizsgázó egy B2 szintű feladatsorhoz tartozó két feladatmegoldása szerepelt. A feladatlapon itt is feltüntettük a feladatleírásokat és az értékelési szempontokat. Értékelést mind a beszédkészség, mind az íráskészség esetén csupán egy alkalommal lehetett beküldeni. Mindkét feladatlap kitöltésére hozzávetőleg 25-25 percet irányoztunk elő, de szigorú időkorlátot nem iktattunk a rendszerbe (a vizsgáztató-továbbképzés további részleteiről lásd Kiszely 2018).

Feladatok és értékelési skálák

A BME általános nyelvvizsgarendszere a vizsgázók beszédkészségét B2 szinten három feladat segítségével méri. Az elsőben a vizsgázónak néhány kifejtő kérdés alapján kell a mindennapi életben előforduló témák széles körében saját személyével kapcsolatban beszélgetnie a vizsgáztatóval. A második részben a vizsgázó önállóan fejti ki gondolatait egy adott témáról, amelyhez képi stimulus tartozik. A harmadik részben a vizsgázónak szerepjátékot kell eljátszania a vizsgáztatóval, célnyelven leírt szituációs feladat alapján.

A feladatokat analitikus értékelési skála segítségével értékeljük, amely a következő öt kritériumot tartalmazza: *feladatmegoldás, kifejezőkészség, nyelvhelyesség, folyamatosság és koherencia,* valamint *kiejtés*. A *feladatmegoldás* kritérium a feladat végrehajtását, a beszédszándékok megvalósítását, a tartalmi relevanciát, a nyelvi funkciók megfelelő alkalmazását és a beszédértést jelenti. A *kifejezőkészség* a szókincs terjedelmére és alkalmazására, a stílus és a regiszter használatára utal. A *nyelvhelyesség* a morfológiai és szintaktikai elemek változatosságát és alkalmazását jelenti. A *folyamatosság és koherencia* a folyamatosságot és az információ szerkesztettségét, míg a *kiejtés* a hangképzést, a szó-, mondat- és beszédhangsúlyt takarja.

A *feladatmegoldást*, a *kifejezőkészséget* és a *nyelvhelyességet* a három feladat esetén külön-külön értékeljük egy 0-tól 5 pontig terjedő skálán. A *folyamatosság és koherencia* kritériumot az első és a harmadik feladat esetében együttesen használjuk szintén egy 0-tól 5 pontig tartó skálán, mert mindkét feladat interaktív jellegű dialógus, míg a

második feladat esetén külön értékeljük, mert ez egy monológ. A *kiejtés* kritériumot a három feladaton átívelően egyszer értékeljük szintén egy 0-tól 5 pontig tartó skálán. Az egyes szempontokra 0–5 pont adható, így összesen 60 nyerspontot lehet elérni.

A BME általános nyelvvizsgarendszere a vizsgázók íráskészségét B2 szinten két feladat segítségével méri. Az elsőben a vizsgázónak informális e-mailt kell írnia egy külföldi ismerősnek általános témákban 170–200 szóban négy megadott tartalmi pont alapján. A második részben internetes fórumhozzászólást kell írni 120–140 szóban, szintén négy szempont alapján.

A két feladatot az értékelők külön-külön értékelik analitikus értékelési skála segítségével, amely a következő négy kritériumot tartalmazza: *kommunikatív érték*, *kifejezőkészség*, *nyelvhelyesség* és *általános benyomás*. A *kommunikatív érték* kritérium a kommunikációs cél elérését, a tartalmi pontok kidolgozását, a kohéziót és a szöveg terjedelmét jelenti. Ez a kritérium gyakorlatilag megfeleltethető a beszédkészségnél alkalmazott *feladatmegoldás* szempontnak. A *kifejezőkészség* és a *nyelvhelyesség* kritériumok ugyanazt jelentik, mint a beszédkészség esetén, amelyekhez természetesen a helyesírás még hozzáadódik. Az *általános benyomás* kritérium alatt a szöveg természetességét, illetve az olvasóra gyakorolt hatását értjük. Az egyes kritériumokra 0–5 pont adható, így összesen 40 nyerspontot lehet szerezni.

Az elemzés eszközei

A kutatás a klasszikus és a modern (probabilisztikus) tesztelmélet alapján készült programok segítségével kereste a kérdésekre a választ. A klasszikus tesztelméleten alapuló leíró statisztikai adatokat az Excel táblázatkezelő, az összefüggés- és különbözőségvizsgálatok adatait pedig az SPSS-programcsomag segítségével számoltuk ki.

A modern tesztelmélet alapján kimutatható eredményeket a Facets programmal (Linacre 2018) állítottuk elő. Ez utóbbi mind az értékelőkről, mind a skáláról, mind pedig a vizsgázókról olyan információkat képes nyújtani, amire a klasszikus tesztelmélet nem képes (Bond–Fox 2015). Facetsnek az értékelést befolyásoló tényezőket nevezzük; ilyen például a vizsgázói nyelvtudás, az értékelői szigor, az értékelési skála kritériumai, a téma nehézsége és maga a feladat is, hogy csak néhányat említsünk a sok közül. A Facets program elvi alapját jelentő többtényezős Rasch-modell (*Many-Facet Rasch Measurement*, rövidítése: MFRM) képes megállapítani a vizsgázók vizsgán elért konkrét eredményét, a vizsgázók becsült képességszintjét, az értékelési szempontok megfigyelt és becsült nehézségét és az értékelők megfigyelt és becsült szigorát azzal a céllal, hogy semlegesítse az értékelői és egyéb hatásokat, azaz a konstruktum-irreleváns tényezőket a vizsgázók teljesítményének értékelésében. Az MFRM ugyanarra a skálára helyezi az értékelői szigorúságot, az itemek (kritériumok) nehézségét és a vizsgázók képességszintjét (Bond–Fox 2015). Más szavakkal kifejezve tehát az egyes kritériumokra adott pontok alapján a többtényezős Rasch-modell egy

valószínűségi becslést, egy modellt ad arra vonatkozóan, hogy egy bizonyos értékelő egy bizonyos vizsgázói teljesítmény egy bizonyos kritériumára milyen valószínűséggel ad egy bizonyos pontszámot, és ezt a becslést összehasonlítja a tényleges pontszámokkal (McNamara 1996). (A Rasch-modellről magyarul lásd még pl. Molnár 2013.)

Eredmények és elemzés

Az alábbiakban a kutatási kérdések sorrendjében közöljük és elemezzük az adatokat.

1. Milyen jellemzőkkel rendelkeznek a BME B2 szintű beszédkészség és az íráskészség értékelési skálái a vizsgáztatói értékelések tükrében?

Az *1.* és a *2. táblázat* a Facets program segítségével készült. Ezeken a két év adatai készségenként és kritériumonként egyben láthatóak, például az 1. táblázat első sorában a 2. feladat nyelvhelyesség *(nyelvhelyesség2)* kritériumának 2017-es és 2018-as adatai összesítve szerepelnek.

Az első oszlopban a kritériumra adott összpontszám szerepel. A második oszlop az értékelések számát jelenti, ami azért 134, mert a 67 vizsgáztató az adott kritériumot kétszer értékelte, egyszer az egyik évben, a következő évben pedig még egyszer (67 × 2 = 134). A harmadik oszlop az adott szempontra adott pontszámok átlagát mutatja, míg a negyedik oszlop a program által elvárt (fair) átlageredményt tartalmazza. Ez az a számsor, amit a többtényezős Rasch-modell alapján a Facets megállapít, kizárva, vagy legalábbis csökkentve a nem kívánatos értékelői hatásokat. Az ötödik oszlop a logitban meghatározott nehézségi értéket mutatja: minél nagyobb ez a szám, annál nehezebb az item, azaz annál szigorúbban értékelték azt a kritériumot. Az infit mnsq és az outfit mnsq értékek az úgynevezett illeszkedési (fit) statisztikák, amelyeknek értékelési skálák esetén Wright–Linacre (1994) szerint a 0,6–1,4 sávban kell mozogniuk. Amennyiben ez a helyzet, akkor az értékelők az elvártaknak megfelelően használják a kritériumokat. Az infit és outfit zstd értékek, amelyek statisztikai hipotézis-tesztként funkcionálnak, –2 és +2 között megfelelőek (Linacre 2003). A legutolsó oszlopban a kritériumok találhatóak egymás alatt nehézségi sorrendbe állítva, azaz például a beszédkészség esetén a *nyelvhelyesség2* a legszigorúbban, a *feladatmegoldás1* pedig a legenyhébben értékelt kritérium.

A táblázatok mindkét készségnél azt mutatják, hogy a megfigyelt pontszám minden egyes kritérium esetén, a beszédkészségnél nagyobb, az íráskészségnél kisebb mértékben, magasabb, mint a fair pontszám, tehát enyhébben értékeltek a vizsgáztatók a modell által elvártnál. Ez alól a beszédkészségnél a *feladatmegoldás1*, íráskészségnél pedig az *általános benyomás1* a kivétel, de csupán elhanyagolható mértékben.

A beszédkészség esetén a legszigorúbban a *nyelvhelyesség2,* a *folyékonyság2* és a *nyelvhelyesség3*, legenyhébben pedig a *feladatmegoldás1,* a *feladatmegoldás3*, valamint a *kifejezőkészség1* kritériumokat értékelték. Figyelemre méltó, hogy az 1. feladatot,

azaz az ahhoz tartozó kritériumokat enyhébben értékelték, mint a 2. és 3.-hoz tartozóakat, aminek több oka lehet (átlagok: 1: 3,73; 2: 3,29; 3: 3,55). Az egyik az, hogy valóban itt nyújtották a legjobb teljesítményt a vizsgázók, hiszen talán ez a legkönnyebben teljesíthető feladat, lévén a vizsgázó személyéhez kapcsolódóan folyik a társalgás, míg a 2. és a 3. feladat nagyobb kihívás elé állítja a vizsgázókat. A másik lehetséges ok pedig pszichológiai természetű a vizsgáztatók oldaláról: nem kizárt, hogy az első feladatot még enyhébben értékelik, mivel ez a kezdete a vizsgának, amikor még a vizsgázók kissé feszültebbek, ezért ekkor még nem fog olyan erősen a ceruza. Megjegyzendő azonban, hogy az átlagok közötti különbség egyik esetben sem volt szignifikáns, tehát messzemenő következtetést jelen adatokból levonni nem lehet (Wilcoxon próba: 1–2: Z = -1,604 p = 0,109; 1–3: Z = -0,272 p = 0,785; 2–3: Z = -1,826 p = 0,068).

Az íráskészség esetén a legszigorúbban a *kommunikatív érték2,* a *nyelvhelyesség2* és a *nyelvhelyesség1*, legenyhébben pedig az *általános benyomás1,* a *kommunikatív érték1* és a *kifejezőkészség2* kritériumokat értékelték. Érdekes, hogy az 1. feladatot, azaz az ahhoz tartozó kritériumokat enyhébben értékelték, mint a 2.-hoz tartozóakat (átlagok: 1: 3,66; 2: 3,46). Ennek az lehet az oka, hogy annak ellenére, hogy ez utóbbi feladattípus (internetes fórumhozzászólás) már 2016 novembere óta részét képezi a vizsgafeladatoknak, az értékelők még mindig nem tudnak megfelelő módon különbséget tenni az 1. és 2. feladattal kapcsolatos elvárások között, azaz az első feladatra vonatkozó elvárásaikat rávetítik a másodikra, de ott nem kapják meg ugyanazt a kidolgozottságot, hiszen ez egy más műfajú és rövidebb írásmű, mint az első, ezért szigorúbban értékelnek. Természetesen az is lehetséges, hogy a vizsgázók valóban jobb teljesítményt nyújtottak ezen szempont mentén az 1. feladatban, mint a másodikban. Az átlagok közötti különbség azonban ebben az esetben sem volt szignifikáns, tehát az eredmény a véletlen műve is lehet (Wilcoxon próba: Z = -1,095 p = 0,273).

Az illeszkedési (infit és outfit) statisztikák egy-egy kivétellel az elvárt 0,6–1,4 értékek között mozognak mindkét készség szinte minden kritériuma esetén, ami azt jelenti, hogy készségenként az összes kritérium egységesen hozzájárul a készség egydimenziósságának bizonyításához, azaz a készségekhez tartozó kritériumok ugyanannak a készségnek egy-egy részképessége, ugyanannak a készségnek az alkomponensei (Eckes 2009). Az egyik kivétel a beszédkészségskálán található *kiejtés* kritérium, ahol az illeszkedési statisztikák túl alacsonyak (0,47 és 0,42) csakúgy, mint a zstd értékek, ami, bár nem teljesen egyértelműen, utalhat arra, hogy ezen kritérium esetén csoportszinten egy jellegzetes értékelői hatásról, a szakirodalmi áttekintésben említett középső pontszámra törekvés tendenciáról beszéljünk. Utóbbira a többi kritérium esetén nincs utalás az adatok alapján. Az íráskészség esetén a kivétel a *nyelvhelyesség1* kritérium, ahol mind az infit, mind az outfit értékek kismértékben ugyan, de magasnak bizonyultak (1,43 és 1,43), és a zstd értékek itt sem megfelelőek, ami elvben azt jelentené, hogy ennél a kritériumnál tapasztalható a legnagyobb variabilitás, azaz

kissé zűrzavaros az értékelés. Ennek azonban esetünkben nincs jelentősége a kritériumhasználatot tekintve, ugyanis a magas fit értékek oka bizonyára az, hogy ez az a kritérium, amelynél a legnagyobb a különbség a 2017-ben és a 2018-ban adott pontszámok között (lásd 7. és 8. táblázat: 2,7015 vs 4,1791) és emiatt tűnik úgy, hogy túlságosan széttartanak az értékelők értékelései.

A táblázatok alsó soraiban található szeparációs megbízhatósági index mindkét esetben magas, a maximális 1-hez közelít (0,97 és 0,93), ami azt jelenti, hogy az értékelők megbízhatóan képesek különbséget tenni a kritériumok között, azaz csoportszintű holdudvarhatásról nem beszélhetünk (Myford–Wolfe 2004).

Össz. erd.	Összes kitölt.	Megf. átlag	Fair(M) átlag	- Logit	Mérési hiba	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Becs. Diszk.	Korreláció Tényl.	Korreláció Elvá.	No.Kritérium
428	134	3.19	3.12	1.31	.17	.89	-1.0	.92	-.6	1.13	.77	.78	6 Nyelvhely.2
430	134	3.21	3.13	1.25	.17	.92	-.6	.91	-.8	1.08	.79	.78	7 Foly.2
439	134	3.28	3.19	.98	.18	1.07	.6	1.07	.6	.94	.83	.79	10 Nyelvhely.3
452	134	3.37	3.28	.58	.18	1.28	2.0	1.30	2.1	.70	.70	.79	4 Feladatm.2
453	134	3.38	3.29	.54	.18	1.14	1.1	1.17	1.2	.82	.71	.79	5 Kif.készség.2
454	134	3.39	3.30	.51	.18	1.04	.3	1.06	.4	.95	.77	.79	3 Nyelvhely.1
458	134	3.42	3.33	.39	.18	1.15	1.1	1.17	1.2	.85	.79	.80	9 Kif.készség.3
472	134	3.52	3.45	-.07	.18	.47	-4.8	.42	-5.2	1.48	.87	.80	12 Kiejtés
494	134	3.69	3.66	-.79	.18	.98	-.1	.96	-.2	1.03	.86	.81	11 Foly.1-3
497	134	3.71	3.68	-.89	.18	.86	-1.0	.87	-.9	1.12	.87	.81	2 Kif.készség.1
508	134	3.79	3.78	-1.25	.18	1.11	.9	1.17	1.2	.83	.81	.81	8 Feladatm.3
548	134	4.09	4.10	-2.56	.18	.90	-.8	.82	-1.2	1.15	.83	.81	1 Feladatm.1
469.4	134.0	3.50	3.44	.00	.18	.98	-.2	.99	-.2		.80		Mean (Count: 12)
34.4	.0	.26	.29	1.11	.00	.20	1.7	.22	1.8		.05		S.D. (Population)
35.9	.0	.27	.30	1.16	.00	.20	1.8	.23	1.9		.06		S.D. (Sample)

Model, Populn: RMSE .18 Adj (True) S.D. 1.10 Szeparáció 6.15 Rétegek 8.53 Megbízhatóság .97
Model, Sample: RMSE .18 Adj (True) S.D. 1.15 Szeparáció 6.43 Rétegek 8.90 Megbízhatóság .98
Model, Fixed (all same) chi-square: 457.8 d.f.: 11 szignifikancia (valószínűség): .00
Model, Random (normal) chi-square: 10.7 d.f.: 10 szignifikancia (valószínűség): .38

***1. táblázat.* Kritériumok elemzése: beszédkészség 2017–2018**

Össz. ered.	Összes kitölt.	Megf. átlag	Fair(M) átlag	- Logit	Mérési hiba	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Becs. Diszk.	Korreláció Tényl.	Korreláció Elvá.	No.Kritérium
429	134	3.20	3.18	1.14	.16	1.21	1.6	1.21	1.6	.80	.47	.62	5 Komm.érték2
460	134	3.43	3.41	.40	.15	.87	-1.0	.86	-1.1	1.12	.63	.63	7 Nyelvhely.2
461	134	3.44	3.41	.37	.15	1.43	3.0	1.43	3.0	.57	.69	.63	3 Nyelvhely.1
474	134	3.54	3.52	.07	.15	.89	-.9	.88	-.9	1.12	.69	.64	8 Alt.benyomás2
489	134	3.65	3.64	-.28	.15	.71	-2.6	.71	-2.6	1.31	.73	.64	2 Kif.készség1
492	134	3.67	3.66	-.35	.15	.88	-1.0	.88	-.9	1.11	.59	.64	6 Kif.készség2
498	134	3.72	3.71	-.49	.15	1.14	1.1	1.16	1.3	.82	.47	.64	1 Komm.érték1
513	134	3.83	3.83	-.84	.15	.86	-1.2	.84	-1.4	1.21	.78	.64	4 Alt.benyomás1
477.0	134.0	3.56	3.54	.00	.15	1.00	-.1	1.00	-.1		.63		Mean (Count: 8)
24.9	.0	.19	.19	.59	.00	.22	1.8	.23	1.8		.11		S.D. (Population)
26.6	.0	.20	.21	.63	.00	.24	1.9	.24	1.9		.11		S.D. (Sample)

Model, Populn: RMSE .15 Adj (True) S.D. .57 Szeparáció 3.70 Rétegek 5.26 Megbízhatóság.93
Model, Sample: RMSE .15 Adj (True) S.D. .61 Szeparáció 3.97 Rétegek 5.63 Megbízhatóság.94
Model, Fixed (all same) chi-square: 116.6 d.f.: 7 szignifikancia (valószínűség): .00
Model, Random (normal) chi-square: 6.6 d.f.: 6 szignifikancia (valószínűség): .36

***2. táblázat*. Kritériumok elemzése: íráskészség 2017–2018**

2. *Feladatonként a beszédkészség és az íráskészség teljesítményeket mely kritériumok mentén értékelték a legszigorúbban és a legenyhébben a vizsgáztatók 2017-ben és 2018-ban?*

A második kutatási kérdésre a klasszikus tesztelmélet alapján kerestük a választ, ugyanis a kérdés megválaszolásához modern tesztelméleti módszerekre nem volt szükség. A vizsgálatot az Excel táblázatkezelő és az SPSS programcsomag segítségével végeztük.

A *3., 4., 7.* és *8. táblázatokban* az első sorban az év mellett az egyes értékelési szempontok szerepelnek rövidítve (a rövidítések jegyzékét lásd a függelékben). A rövidítések melletti szám az adott feladat sorszámára utal, tehát például a *Kif.2* a 2., azaz az önálló témakifejtés feladat *kifejezőkészség* kritériumát jelenti. A második sor a feladatot értékelő vizsgáztatók számát jelenti (67). A többi sorban a leíró statisztika legfontosabb mutatói találhatóak.

A beszédkészség táblázatokban (*3.* és *4.*) szereplő átlagokból az látszik, hogy amennyiben feladatonként tekintjük a számokat, akkor a két év során a hat feladatból öt alkalommal a *nyelvhelyesség* szempontot értékelték a legszigorúbban a vizsgáztatók, ami azért érdekes mert a két vizsgázói teljesítmény színvonala nagyban különbözött egymástól (2017: 37 vs. 2018: 50 nyerspont a megszerezhető 60-ból), ám ez mégsem változtatott a kritériumok megítélésén. Az elvégzett páros Wilcoxon-próbák (amelynek során a *nyelvhelyesség* és az adott feladathoz tartozó többi kritérium átlagai közötti különbség szignifikanciaszintjét vizsgáltuk meg) azt mutatták, hogy a *nyelvhelyesség* szigorúsága a 13 esetből nyolcszor valós különbséget takar ($p < 0{,}05$), azaz az eredmény nem a véletlen műve (*5.* és *6. táblázat*).

2017	Fel.1	Kif.1	Ny.1	Fel.2	Kif.2	Ny.2	Foly.2	Fel.3	Kif.3	Ny.3	Foly.1-3	Kiej.	Összes
N	67	67	67	67	67	67	67	67	67	67	67	67	67
Átlag	3,462	2,955	2,836	2,880	2,910	2,657	2,642	3,134	2,761	2,582	2,955	2,97	34,75
Módusz	3	3	3	3	3	3	3	3	3	3	3	3	33
Medián	3	3	3	3	3	3	3	3	3	3	3	3	34
Szórás	0,502	0,405	0,539	0,564	0,570	0,509	0,513	0,6	0,525	0,497	0,474	0,171	3,221

***3. táblázat.* Leíró statisztika: beszédkészség 2017**

2018	Fel.1	Kif.1	Ny.1	Fel.2	Kif.2	Ny.2	Foly.2	Fel.3	Kif.3	Ny.3	Foly.1-3	Kiej.	Összes
N	67	67	67	67	67	67	67	67	67	67	67	67	67
Átlag	4,716	4,462	3,940	3,865	3,850	3,731	3,776	4,447	4,074	3,970	4,417	4,074	49,32
Módusz	5	4	4	4	4	4	4	4	4	4	4	4	48
Medián	5	4	4	4	4	4	4	4	4	4	4	4	49
Szórás	0,454	0,531	0,518	0,625	0,500	0,479	0,572	0,530	0,531	0,626	0,554	0,470	4,020

***4. táblázat.* Leíró statisztika: beszédkészség 2018**

2017	Ny.1 – Fel.1	Ny.1 – Kif.1	Kif.3 – Ny.3	Fel.3 – Ny.3	Foly.13 – Ny.3
Z	-6,410[a]	-1,569[b]	-2,191[b]	-5,778[b]	-4,642[b]
Asymp. Sig. (2-tailed)	,000	,117	,028	,000	,000

5. táblázat. **Beszédkészség különbözőségvizsgálat 2017: Wilcoxon-próba**

2018	Ny.1 – Fel.1	Kif.1 – Ny.1	Kif.2 – Ny.2	Fel.2 – Ny.2	Ny.2 – Foly.2	Kif.3 – Ny.3	Fel.3 – Ny.3	Foly.13 – Ny.3
Z	-6,761[a]	-5,840[b]	-1,569[b]	-1,567[b]	-,577[a]	-1,400[b]	-5,154[b]	-4,727[b]
Asymp. Sig. (2-tailed)	,000	,000	,117	,117	,564	,162	,000	,000

6. táblázat. **Beszédkészség különbözőségvizsgálat 2018: Wilcoxon-próba**

Az íráskészség táblázatokban (7. és 8.) szereplő átlagokból kiolvasható, hogy amennyiben feladatonként tekintjük a számokat, akkor a két év során a négy feladatból három alkalommal a *kommunikatív érték* szempontot értékelték a legszigorúbban a vizsgáztatók, ami azért figyelemre méltó, mert a két vizsgázói teljesítmény színvonala nagyban különbözött egymástól (2017: 28 vs. 2018: 34 pont a megszerezhető 40-ből), ám ez mégsem változtatott a kritériumok megítélésén. A *nyelvhelyesség* kritériumot csupán a 2017-es 1. feladat esetén értékelték a legszigorúbban. Megjegyzendő, hogy ebben a feladatban éppen a *kommunikatív érték* lett a legenyhébben értékelt szempont, szöges ellentétben a másik három feladattal. Az itt is elvégzett Wilcoxon-próbák, amelynek során a *kommunikatív érték* és az adott feladathoz tartozó többi kritérium átlagai közötti különbség szignifikanciaszintjét vizsgáltuk meg, azt mutatták, hogy a *kommunikatív érték* szigorúsága a kilenc esetből nyolcszor valós különbséget takar ($p < 0,05$), azaz az eredmény nem a véletlen műve (*9.* és *10. táblázat*).

2017	Komm.ért.1	Kif.1	Nyelvh.1	Ált.b.1	Komm.ért.2	Kif.2	Nyelvh.2	Ált.b.2	Összes
N	67	67	67	67	67	67	67	67	67
Átlag	3,4925	3,1493	2,7015	3,2239	2,9851	3,4776	3,1045	3,2537	25,3881
Módusz	4	3	3	3	3	4	3	3	23
Medián	4	3	3	3	3	4	3	3	25
Szórás	0,5871	0,4353	0,4928	0,4546	0,6852	0,6120	0,5261	0,6116	2,4738

7. táblázat. **Leíró statisztika: íráskészség 2017**

2018	Komm.ért.1	Kif.1	Nyelvh.1	Ált.b.1	Komm.ért.2	Kif.2	Nyelvh.2	Ált.b.2	Összes
N	67	67	67	67	67	67	67	67	67
Átlag	3,9403	4,1493	4,1791	4,4328	3,4179	3,8657	3,7612	3,8209	31,5672
Módusz	4	4	4	5	3	4	4	4	30
Medián	4	4	4	4	3	4	4	4	32
Szórás	0,6715	0,5295	0,5484	0,6086	0,6068	0,6251	0,6534	0,7370	3,0213

8. táblázat. **Leíró statisztika: íráskészség 2018**

2017	Kif.2 – Komm.ért.2	Ny.2 – Komm.ért.2	Ált.b.2 – Komm.ért.2
Z	–4,814[a]	–1,241[a]	–3,000[a]
Asymp. Sig. (2-tailed)	,000	,215	,003

9. táblázat: Íráskészség különbözőségvizsgálat 2017: Wilcoxon-próba

9. táblázat. Íráskészség különbözőségvizsgálat 2017: Wilcoxon-próba

2018	Kif.1 – Komm.ért.1	Ny.1 – Komm.ért.1	Ált.b.1 – Komm.ért.1	Kif.2 – Komm.ért.2	Ny.2 – Komm.ért.2	Ált.b.2 – Komm.ért.2
Z	–2,556[a]	–2,317[a]	–4,907[a]	–4,328[a]	–3,450[a]	–3,430[a]
Asymp. Sig. (2-tailed)	,011	,021	,000	,000	,001	,001

10. táblázat. Íráskészség különbözőségvizsgálat 2018: Wilcoxon-próba

Az egyik legnagyobb különbség tehát a beszédkészség és az íráskészség feladatok értékelése között az, hogy a beszédkészség esetén, hasonlóan McNamara (1990) és Corrigan (2007) eredményeihez, többségében a *nyelvhelyesség* a legszigorúbban értékelt kritérium, az íráskészség esetén pedig a *kommunikatív érték* a domináns, ami tulajdonképpen a beszédkészség skála *feladatmegoldás* kritériumának feleltethető meg. Ennek a magyarázata az lehet, hogy a vizsgáztatók azokon a területeken szigorúbbak, amelyeket a legjobban ismernek, mert itt veszik észre legkönnyebben a hibákat (Mike Linacre, e-mail-kommunikáció, 2019), ami a beszédkészség esetén a *nyelvhelyesség*. Érdekes viszont, hogy az íráskészség értékelésekor, ugyanezen vizsgáztatók esetén, ez a magyarázat nem állja meg a helyét, aminek talán az az oka, hogy annyira részletesen kidolgozott az íráskészség-értékelési útmutató a tartalmi pontok kifejtettségét illetően, ami a *kommunikatív érték* kritérium megítélésének elsődleges összetevője, hogy a vizsgáztatók még szigorúbban, illetve még szabálykövetőbben használják ezt a skálát, mint az általuk jól ismert *nyelvhelyesség* skálát. A jelenség magyarázata lehet továbbá az is, hogy az íráskészség esetén a tartalom szempontjából magasabb elvárásokat támasztanak az értékelők a vizsgázókkal szemben, mint a beszédkészség esetén.

3. Milyen összefüggések és különbségek mutathatók ki a beszédkészség és az íráskészség értékelésére használt skálák alkalmazása között?

A harmadik kérdésre, a másodikhoz hasonlóan, a klasszikus tesztelmélet alapján kerestük a választ, az SPSS programcsomag felhasználásával. A korrelációelemzéssel azt kívántuk megvizsgálni, hogy a vizsgáztatók mennyire azonos tendencia szerint értékelik a beszédkészség- és az íráskészség-teljesítményeket. Mivel a tesztpontszámok nem intervallum, hanem rangsor skálán helyezkednek el, ezért a Spearman-féle rangkorrelációs együtthatót számoltuk ki, amelynek az a lényege, hogy nem az adott pontszámok, hanem a sorban elfoglalt hely szerint állapítja meg az összefüggést két adatsor között.

A beszédkészség és az íráskészség értékelések korrelációs együtthatója negatív előjelű ($r = -0{,}074$, $p = 0{,}554$), ami azt a tendenciát rajzolta ki, hogy a vizsgáztatók alapvetően ellentétesen ítélték meg a beszéd- és írásteljesítményeket, azaz aki az egyiket szigorúbban értékelte, az a másikat engedékenyebben. Ez az együttható azonban nem szignifikáns, ami azt jelenti, hogy az eredmény a véletlen műve is lehet, tehát további vizsgálatok szükségesek annak megállapítására, hogy ez a jelenség milyen mértékben valós. A vizsgált csoportról azonban megerősítette azt a korábbi elemzések által is előrejelzett képet, hogy különbözően használják a két skálát, mások a prioritások egyik és másik esetben.

A beszéd- és az íráskészség adatsorral kapcsolatban különbözőségvizsgálatot is végeztünk azzal a céllal, hogy kiderüljön, hogy a két adatsor, azaz a teljesítményekre adott értékelések átlagai (százalékban kifejezve: beszédkészség: 70,21%; íráskészség: 71,36%) milyen mértékben térnek el egymástól. A páros t-próba az átlagok között nem mutatott ki szignifikáns különbséget, azaz az értékelők átlageredményei nagyon hasonlóak voltak egymáshoz ($t = -1{,}276$, $df = 66$, $p = 0{,}206$).

Az összefüggés- és a különbözőségvizsgálatok tehát nagyon érdekes eredményeket hoztak: a vizsgáztatók ugyan ellentétes módon értékelték a beszédkészség és az íráskészség vizsgázói teljesítményeket, az átlageredményeket tekintve azonban nagyon hasonlóan értékeltek. Vizsgálatunk célját tekintve ez azt jelenti, hogy az értékelők a skálák egyes kritériumait különböző módon, a teljes skálát viszont nagyon hasonló módon használták. Megjegyzendő azonban, hogy mivel az összefüggés és különbözőségvizsgálatok számítása is nyerspontokkal történt, ezért a különbségek nem csupán a skálahasználatra, hanem a vizsgázói teljesítményekre is vonatkozhatnak.

Összefoglalás

Jelen tanulmány arra tett kísérletet, hogy a BME Nyelvvizsgaközpont 2017. és 2018. évi vizsgáztató-továbbképzése során összegyűlt adatok segítségével összehasonlítsa, hogyan használják ugyanazon angol nyelvi vizsgáztatók az íráskészség és a beszédkészség méréséhez készült értékelési skálákat két-két B2 szintű, de különböző nyelvi színvonalú nyelvvizsgázói teljesítmény értékelése során.

Az eredmények azt mutatták, hogy kevés kivétellel minden kritérium esetén, a beszédkészségnél nagyobb, az íráskészségnél kisebb mértékben, a vizsgáztatók engedékenyebben értékeltek a modern tesztelméleti alapon kialakított modell elvárásainál. Mind a beszédkészség, mind pedig az íráskészség esetén az első feladatot, azaz az ahhoz tartozó kritériumokat, enyhébben értékelték, mint a többi feladathoz tartozóakat, bár ez az adat még további megerősítést igényel. Sikerült kimutatni, hogy a skálák egyes kritériumai az adott készség alkomponensei. Összességében az értékelők megbízhatóan tudtak különbséget tenni az értékelési szempontok között, azaz csoportszinten holdudvarhatásról vagy középső pontszámra törekvés tendenciáról nem beszélhetünk.

Feladatonként tekintve a kritériumokkal szembeni szigorúság kérdésében érdekes megállapítást tehettünk. A beszédkészség esetén a *nyelvhelyesség*, az íráskészség esetén pedig a *kommunikatív érték* bizonyult a legszigorúbban értékelt kritériumnak. A *nyelvhelyesség* primátusát azzal magyaráztuk, hogy a vizsgáztatók azokon a területeken szigorúbbak, amelyeket a legjobban ismernek, az íráskészség esetén azonban a *kommunikatív érték* dominanciája felülírta ezt a szabályt, amit az értékelési útmutató rendkívüli részletezettségével, illetve a tartalommal kapcsolatos elvárások különbözőségével magyaráztunk. További kutatások szükségesek annak feltárására, hogy ez a jelenség általánosítható-e. Érdemes azt is megvizsgálni, hogy más nyelvek esetében hasonló eredmények születnének-e.

A vizsgálatok során az is kiderült, hogy a vizsgált értékelői csoport ugyan ellentétes módon értékelte a beszédkészség- és az íráskészség-vizsgázói teljesítményeket, azaz aki az egyiket szigorúbban értékelte, az a másikat engedékenyebben, az átlageredményeket tekintve azonban nagyon hasonlóan. Vizsgálatunk fókuszára nézve ez azt jelenti, hogy az értékelők a skálák egyes kritériumait különböző módon, a teljes skálát viszont nagyon hasonló módon használták.

A kutatás eredményei fontos információkat nyújthatnak a vizsgaközpont számára. A beszédkészségre koncentráló vizsgáztató-továbbképzéseken az első feladat a másik kettőhöz képest történő enyhébb megítélésére érdemes fókuszálni. Ki lehet térni továbbá a *feladatmegoldás* és a *nyelvhelyesség* szempontok megítélésére is, az egyiknél ugyanis a túlzott engedékenység, a másiknál pedig az esetleges túlzott szigor jelenthet problémát. Elképzelhető az is, hogy az értékelési kritériumokon kell további finomításokat végezni ezen két kritérium esetén.

Az íráskészség esetén a jövőbeni továbbképzéseken azt érdemes megvizsgálni, hogy miért van különbség az első és a második feladat megítélésében, hiszen kiderült, hogy a vizsgázók teljesítményét a viszonylag új második feladaton szigorúbban ítélik meg a vizsgáztatók, mint az első feladat esetében. Ezen készség esetén is elképzelhető, hogy az értékelési skálát felülvizsgálat alá kell vetni, mert lehetséges, hogy nem egyformán alkalmas mindkét feladat értékelésére.

Az eredmények értelmezésekor két körülményt nem szabad elfelejteni. Az egyik, hogy az elemzett értékelést nem valódi vizsgán, hanem egy továbbképzés keretében végezték a résztvevők, amelyek tétje nem hasonlítható egymáshoz. A másik pedig az, hogy a vizsgálat online válaszadáson alapult, ami elképzelhető, hogy gondot okozott a papír-toll alapú vizsgáztatáshoz szokott vizsgáztatók számára (Myford–Wolfe 2003). A vizsgálatokat a fent említett irányok mellett vizsgáztatókkal készült interjúkkal és hangos gondolkodtatásos (*think-aloud protocol*) adatfelvételi eljárással is érdemes kiegészíteni.

IRODALOM

Benke Eszter (2008): Rater and rating scale interaction in the validation of the assessment of writing performance. In: Majoros Pál (szerk.) *BGF Tudományos Évkönyv 2007: Reformok útján*. Budapest: Budapesti Gazdasági Főiskola, 327–335.

Bond, T. – Fox, C. (2015): *Applying the Rasch model: Fundamental measurement in the human sciences.* Routledge: New York and London.

Corrigan, M. (2007): *Seminar to calibrate examples of spoken performance: Report on the analysis of rating data.* Strasbourg, France: Council of Europe/Language Policy Division.

Eckes, T. (2015): *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments.* Frankfurt am Main: Peter Lang.

Eckes, T. (2019): Many-facet Rasch measurement: Implications for rater-mediated language assessment. In: Aryadoust, V. – Raquel, M. (eds.) *Quantitative data analysis for language assessment: Volume I: Fundamental techniques.* London and New York: Routledge, 153–175.

Eckes, T. (2009): Many-facet Rasch measurement. In: Takala, S. (ed.) *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (Section H).* Strasbourg, France: Council of Europe/Language Policy Division.

Kiszely Zoltán (2018): Egy online vizsgáztató-továbbképzés tapasztalatai. *Modern Nyelvoktatás,* 24/1, 3–18.

Kuiken, F. – Vedder, I. (2014): Raters' decisions, rating procedures and rating scales. *Language Testing*, 31/3, 279–284.

Linacre, J. M. (2003): Rasch Power Analysis: Size vs. Significance: Standardized Chi-Square Fit Statistic. *Rasch Measurement Transactions*, 17/1, 918, https://www.rasch.org/rmt/rmt171n.htm

Linacre, J. M. (2018): *A user's guide to FACETS: Rasch-model computer programs. Program manual.* Chicago, IL: Winsteps.com. Letöltve www.winsteps.com/facets.htm

Lumley, T. – McNamara, T. F. (1995): Rater characteristics and rater bias: Implications for training. *Language Testing*, 12/1, 54–71.

McNamara, T. (1990): Item response theory and the validation of an ESP test for health professionals. *Language Testing,* 7/1, 52–75.

McNamara, T. (1996): *Measuring second language performance*. Longman: Harlow.

Molnár Gyöngyvér (2013): *A Rasch-modell alkalmazási lehetőségei az empirikus kutatások gyakorlatában.* Gondolat: Budapest.

Myford, C. M. – Wolfe, E. W. (2003): Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, 4, 386–422.

Myford, C. M. – Wolfe, E. W. (2004): Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, 5, 189–227.

Wind, S. A. – Peterson, M. E. (2017): A systematic review of methods for evaluating rating quality in language assessment. *Language Testing,* 35/2, 161–192.

Wright, B. D. – Linacre, J. M. (1994): Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8/3, 369–370. https://rasch.org/rmt/rmt83b.htm

FÜGGELÉK

Rövidítések jegyzéke:

Fel.1	Feladatmegoldás 1
Kif.1	Kifejezőkészség 1
Ny.1	Nyelvhelyesség 1
Fel.2	Feladatmegoldás 2
Kif.2	Kifejezőkészség 2
Ny.2	Nyelvhelyesség 2
Foly.2	Folyamatosság és koherencia 2
Fel.3	Feladatmegoldás 3
Kif.3	Kifejezőkészség 3
Ny.3	Nyelvhelyesség 3

Foly.1-3	Folyamatosság és koherencia 1, 3
Kiej.	Kiejtés
Komm.ért.1	Kommunikatív érték 1
Kif.1	Kifejezőkészség 1
Nyelvh.1	Nyelvhelyesség 1
Ált.b.1	Általános benyomás 1
Komm.ért.2	Kommunikatív érték 2
Kif.12	Kifejezőkészség 2
Nyelvh.2	Nyelvhelyesség 2
Ált.b.2	Általános benyomás 2