 https://doi.org/10.14232/iskkult.2026.5-6.23

Dávid Gergely András

ELTE Angol-Amerikai Intézet Angol Nyelvpedagógia Tanszék

A kettős párhuzamos értékelés az angoltanári szakdolgozatok értékelésében

Az érvényesítés nem csupán elméleti teljesítmény, hanem értékítéletek, eredmények, érdemjegyek és pontszámok értelmezhetősége és hitelessége megállapításának céljával végrehajtott adatgyűjtés és elemzés, a statisztikai megbízhatóság kimutatásával.

A kutatás témája az ELTE Angol-Amerikai Intézete végzős, az angol nyelv és kultúra tanára szakos hallgatók (osztatlan tanárképzés, OTAK) szakdolgozatainak (a továbbiakban angoltanári szakdolgozatok) kettős párhuzamos értékelése, a 2011 óta változatlan formában alkalmazott analitikus skálák megfelelése (beválása) alapján.

A szerző hite szerint bármilyen mérési-értékelési rendszert – különösen az érvényesítés (validálás) előfeltételeként – időről-időre vizsgálni kell, hogy a mérés eszközei mennyiben váltják be a hozzájuk fűzhető, jellemző szakmai elvárásokat (megfelelés, beválás), vagyis a mérés konzisztenciáját, megbízhatóságát. A tanulmány legfőbb célja az angoltanári szakdolgozatok kettős értékelésének mérlegre tétele, tekintettel arra, hogy a kettős értékelés módszerének megbízhatóságát, értelmét egyes források (pl. Meadows és Billington, 2005) a szükséges ráfordítások okán időről időre vitatják.

A tanulmányban sor kerül a kettős értékelés itt használt fogalmának meghatározására, mert míg az angol nyelvű irodalomban a *kettős értékelés* kifejezés alatt általában két értékelő egyidejű, párhuzamos pontozását értik, az olvasói reflexiók alapján világos – figyelembe véve az angoltanári szakdolgozatok értékelésének magyarországi gyakorlatát is –, hogy e tanulmányban részletesebben kell leírni, hogyan értendő a tárgyalt kettős értékelés.

A megbízhatóságot érintő kötelező kérdések megválaszolását követi az értékelés eredményének (az osztályzatoknak) a vizsgálata, vagyis e kutatás így lépéseket tesz a szakdolgozati érdemjegyek és a kapcsolódó döntések (érdemjegy, siker vagy bukás) érvényességének (validitásának) megállapítása felé.

A skálák és értékelők mint az értékelés eredményét befolyásoló legfőbb performancia (nyelvhasználat, megvalósuló nyelvhasználat) tényezőinek bemutatása mellett a tanulmány célja az is, hogy rámutasson, az értékelői hatás nem homogén tényező. Rámutat specifikusan arra, hogy az értékelői hatás a szakdolgozati bírálók és témavezetők szerepei mentén eltérhet a kettős értékelés keretén belül. A cikkben továbbá bemutatásra kerülnek az értékelői *szerep* tényező létezésének bizonyítékai. Az értékelők az eljárásban betöltött szerepük szerint témavezetők vagy bírálók. Amennyiben a tárgyalás szerepeik szerint nem specifikus, mindkettőt értékelőnek nevezem.

Irodalom

A szakirodalmat az alábbi kérdéskörökre bontva, a terjedelmi korlát miatt szelektíve tárgyalom: Előbb a megbízhatóságnak a skálák megtervezésével, szerkesztésével kapcsolatos aspektusait és a kapcsolódó szkepticizmust tárgyalom, majd rátérek a megbízhatóság kettős értékeléssel kapcsolatos aspektusaira, ez után a skálák és értékelők mint a két legfontosabb performanciahatás következik, végül a mérés és érvényesítés elméleti háttere.

A megbízható értékeléssel kapcsolatos kételyek

A szakdolgozati skálák fejlesztését eredetileg a szakirodalomban felbukkanó, az értékelés megbízhatóságával kapcsolatos, máig ható szkeptikus megfogalmazások kényszerítették ki (Dávid & Piniel, 2018). Néhány korai szerző inkább szerkezeti, tapasztalati, mint empirikus alapon fejezte ki kételyeit a skálák egyes sávjainak (fokozatok) lehetséges legmagasabb számát vizsgálva: Az értékelés megbízhatósága forog kockán, ha a sávok (fokozatok) száma nagyobb, mint öt (Pollitt, 1991. 90.) vagy hét (Alderson és mtsai, 1995. 111.). Luoma (2004. 80.) hasonló értelemben ír, hozzátéve, hogy a vizsgarendszerekben alkalmazott skálák sávjainak jellemző száma 4-6. Weigle (2002. 123.) nem számszerűsíti, csak sugallja, hogy a sávok számának felső ésszerű határa van.

A szkepszis érinti az értékelési szempontok (skálák, kategóriák) legmagasabb számát is. A Közös Európai Referenciakeret (KER) szerint négy vagy öt értékelési szempont már erős kognitív terhelés, hét szempont pedig már a pszichológiai felső határ (PTMIK, 2002. 239.), míg Luoma (2004. 80.) szerint 5-6 értékelési szempont majdnem a maximum. Ezzel együtt feltűnik, hogy néhány más skála jóval több kategóriát fog át, mint amit a fentiek alapján várhatnánk. Felvetődik, hogy ha 5-6 értékelési szempont, illetve a nagyszámú sáv a legtöbb, amire az értékelő megfelelő módon figyelni tud, hogyan lehetséges 25 állítást nyomon követni (ld. pl. Yanosky és mtsai, 2013), ha minden állításhoz négyfokozatú skála kapcsolódik.

A kettős értékelés távlatai

A kettős értékelés kérdése szorosan kapcsolódik a megbízhatóság követelményéhez. Meadows és Billington (2005) még azt az állapotot dokumentálja, amikor a kettős értékeléstől a megbízhatóság jelentős növekedése volt várható korábbi kutatások alapján. Partington (1994) még arra a következtetésre jutott, hogy a kettős vagy többes értékelés gyakorlata elterjedőben van. Érzékelhető ugyanakkor, hogy míg a kettős értékelés egyben a munkaterhelés növelését jelenti, szervezési, logisztikai és erőforrás-problémák újabban sem voltak megkerülhetők (Brooks, 2004). Átfogó hivatalos jelentések is jelzik, hogy az értékelés (*impression marking*) megbízhatósága továbbra is kérdés a felmerülő logisztikai problémák mellett. Az Ofqual (2014a) azt is hangsúlyozza, hogy a kettős értékelés csak kis mértékben emeli a mérés megbízhatóságát, míg az Ofqual (2014b) szerint a kettős értékelés, bár jelen van több országban is, Nagy-Britanniában, illetve jórészt a Brit Nemzetközösséghez tartozó országokban a közoktatás terén aligha megvalósítható a nagy vizsgázói létszámok miatt.

A legújabb szakirodalmi kritika szerint kérdéses, hogy a kettős értékelés mennyiben javítja a megbízhatóságot. Steele és Shaw (2022) a két értékelő nyerspontjait vetette össze a klasszikus tesztelmélet szerint. Az eredmény gyenge volt, alacsony egyetértési mutatókkal. Véleményük szerint a modern (probabilisztikus) tesztelmélet (pl. *Item*

Response Theory, vagy a Rasch-módszer) alapján végzett feldolgozással viszont sokkal magasabb egyetértési mutatót kaphatnak. Williams és Kemp (2018) vizsgálatukban oda jutottak, hogy a szakdolgozatok kettős értékelésében alacsony a belső és külső, egymástól független értékelők közötti korreláció. Megállapítják, hogy hozzávetőlegesen olyan alacsony korrelációról (0,39) van szó, mint amilyen pl. szakfolyóiratok lektorainak egyetértésénél tapasztalható. Ezt megerősíti Bramley és Dhawan (2011), akik szerint az értékelők szigora szerinti különbségek a ponteredmények szórásának csak kis részét, a legkisebbet képezik, viszont – vélekedésük szerint – a vizsgázók az egymásnak ellentmondó értékeléseket tűrik a legkevésbé, vagyis alacsony a kettős értékelés felszíni érvényessége (*face validity*): Ha tehát a pontok szórása kicsi, és alacsony a kettős értékelés felszíni érvényessége, az értékelés költségei viszont tetemesek, vissza lehet térni az egyszeres értékeléshez.

A 2010–2011 során kidolgozott angoltanári szakdolgozati értékelési rendszerben (skálák és az eljárás), abban, hogy kettős értékelés valósult meg minden dolgozat esetében, leginkább Alderson és munkatársai (1995. 132–133.) kézikönyvének, illetve az 1990-es évek elején folytatott konzulensi tevékenységüknek a hatása, iránymutatása mutatható ki. A szerzők rutinszerű kettős értékelést, vagyis minden dolgozat értékelése esetében rendszeresen két értékelő alkalmazását javasolják úgy, hogy a két értékelő párhuzamosan, egymástól függetlenül értékel, nem látva egymás pontozását. E hatás nem elválasztható a fent áttekintett angol nyelvű szakirodalmi törekvésektől és szakmai meggyőződésektől, amelyek az elmúlt évtizedeket jellemezték.

A legfontosabb performanciatényezők

A szakdolgozatok esetében – miként a szubjektív értékelésnél általában – feltételezhető, hogy az eredményre számos performanciatényező hat. A skálák és az értékelők azért tekinthetők performanciatényezőknek, mert egyaránt a mérés-értékelés eszközei, szisztematikus változók, de mégsem tekinthetők a mérni kívánt dolgozatírói kompetencia, a mérendő részének. E tényezőkről tudományos úton (és tapasztalati úton is), továbbá a velük járó szórás nagyságát mérve tudjuk, hogy jelentős mértékben befolyásolják az értékelés eredményét, jelen esetben a pontokat és osztályzatokat. McNamara (1996) jó képet ad az idegennyelv-tudás mérésének kommunikatív korszakában végbement elméleti változásokról, amelyek a kompetencia fogalom kibővülését és a performancia fogalmának differenciálódását eredményezték. Alább a skálahatást és az értékelői hatást tekintem át.

Skálák vagy listák az értékelésben?

Az angoltanári szakdolgozati skálák klasszikus (táblázatos) elrendezésű analitikus skálák, melyekben az oszlopok az értékelés hét szempontjának, míg a 0–4 ponttal értékelt sávok a vizsgált készség szintjeinek feleltethetők meg. A pontosság kedvéért hozzá kell tenni, hogy az angol nyelvű szakirodalom használja a *rubrics* szót is az értékelési útmutatók jelölésére, ezt azonban a KER (Council of Europe, 2001) egyáltalán nem használja. Helyette a scale terminust használja, véleményem szerint szerencsésebb elnevezéssel, mert tipológiai alternatíváról van szó.

Lukácsi (2021) a KER-ből kiindulva (PTMIK, 2002. 233–234.) kifejti, hogy az analitikus skálák helyett listák (*checklists*) alkalmazása reális alternatíva. Listák használata esetén az értékelő jellemzően táblázatos elrendezésű szempontok és deskriptorok (kritériumleírások) helyett egy listát tanulmányoz, ahol minden állítás a követelmények egy-egy

eleme. Lukácsi sikeres példákat is hoz, pl. Kim (2011). A KER szerint különösen akkor tekinthetők a listák a skálák alternatívájának, ha a kérdés az, hogy adott követelmények tekintetében adott személy teljesítménye megfelel-e a listában felsorolt követelményeknek. Ha az olvasó szereti a metaforákat, a KER megfogalmazása szerinti „vízszintes orientáció" hasznos lehet (PTMIK, 2002. 234.): A lista tételei szintkövetelményként értelmezhetők, amelyeknek a teljesítmény vagy megfelel, vagy nem. Az ELTE program kontextusában ez a megfelel/nem felel meg (bináris) osztályzást jelentené.

A listás megközelítés mellett a másik megközelítésmód továbbra is a skálák alkalmazása – a klasszikusnak tekinthető megoldás –, amelynek segítségével adott teljesítmény a skála (skálák) valamely sávjában helyezhető el, szöveges kritériumleírással együtt. Ismét egy metafora: A skálás megközelítés a KER szerint „a függőleges tengely menti megoszlás" (PTMIK, 2002. 233.), ahol az értékelés a pontozó ítéletétől függően felfelé vagy lefelé mozog a skálán – diagnosztikus eszköz. A skálás megoldás közelebb áll a szakdolgozatok esetében az ELTE-n alkalmazott osztályzási rendszerhez (1–5), mint a listás megoldás.

Az értékelői hatás számos kutatás középpontjában áll, amit itt csak jelzésszerűen, annak inkább irányait kijelölve lehet bemutatni. Wigglesworth (1993), Kondo-Brown (2002) és Schaefer (2008) az értékelői részrehajlást vizsgálta, míg Winke és munkatársai (2012) az értékelő célnyelvi kiejtési problémáinak ismeretét (ti. az értékelő tanulta-e a vizsgázó anyanyelvét) mint a részrehajlás forrását. Brown (2003) a számítógépen írt vizsgázói válaszokat (szövegeket) a papíralapú vizsgák kézzel írott szövegeinek olvashatóságával vetette össze. Míg Kim (2015) az értékelőkön belül azok tapasztalati eltéréseit és eltérő önfejlesztési sebességét vizsgálta, addig Lamprianou és munkatársai (2023) szerint a tapasztalatok mennyisége mellett azok különféle fajtái is lehetnek olyan tényezők, amelyek az értékelők szigorát és az értékelés megbízhatóságát befolyásolják.

Listás, skálás és vegyes megoldások

Bár a listás megközelítés fő vonzereje, hogy a lista végigvezeti az értékelőt a szempontokon, ahol egyszerű eldöntendő kérdéseket kell megválaszolni, vonzóvá teszi az is, hogy az idegennyelv-tudás mérésének szakterülete mellett a listás megoldás a pszichológia, egészségügy (fogyatékosság) (pl. Sideridis és Padeliadu, 2011), oktatás-menedzsment (pl. Hallinger, 2011) terén is megtalálható. A szakirodalom áttekintése során azonban feltűnő, hogy kevés további tisztán listás megoldást lehet találni. Leszámítva Safari és Ahmadi (2023) munkáját, Lukácsi forrásai a korábbi években keletkeztek. A listás és skálás megközelítés tehát elméletileg vonzó dichotómia, ideértve a vízszintes és függőleges metaforákat, azonban a gyakorlat nem mutat rá éles különbségre. A kutatási beszámolók egy jelentős része egy harmadik, vegyes kategória létezéséről tanúskodik, amelyben az állítások (vagy kérdések) sora a listás megoldást jelzi, viszont mindegyikhez skála kapcsolódik: Minden állításnak saját skálája van, pl. Martensnél (1979), ahol az állításokhoz kapcsolódó skálák nem kevesebb, mint kilenc fokozatúak. A KER a vegyes megoldást nem tekinti külön kategóriának

– nem is említi –, hanem besorolja a listás megoldások közé (PTMIK, 2002. 234.), egyben gyengítve ezzel a skálás és listás megkülönböztetés erejét.

Az értékelők hatása

Az értékelői tevékenység a mérés eredményére ható, talán legfontosabb itt tárgyalt performanciahatás (-tényező). A kutatások jellemző pedagógiai célja az eredményekre gyakorolt értékelői hatás minimalizálása (pl. Kim, 2015. 239.) – nem véletlenül, hiszen a mérési cél (kompetencia) mellett jelentkező performanciatényezők gyakran nem kívánatos módszertényezők: A skálák hatása mellett az értékelő is nem kívánatos hatást fejt ki, hiszen szigorával-elnézőségével a feltételezett kompetenciához képest módosítja a ponteredményeket.

Az értékelői hatás számos kutatás középpontjában áll, amit itt csak jelzésszerűen, annak inkább irányait kijelölve lehet bemutatni. Wigglesworth (1993), Kondo-Brown (2002) és Schaefer (2008) az értékelői részrehajlást vizsgálta, míg Winke és munkatársai (2012) az értékelő célnyelvi kiejtési problémáinak ismeretét (ti. az értékelő tanulta-e a vizsgázó anyanyelvét) mint a részrehajlás forrását. Brown (2003) a számítógépen írt vizsgázói válaszokat (szövegeket) a papíralapú vizsgák kézzel írott szövegeinek olvashatóságával vetette össze. Míg Kim (2015) az értékelőkön belül azok tapasztalati eltéréseit és eltérő önfejlesztési sebességét vizsgálta, addig Lamprianou és munkatársai (2023) szerint a tapasztalatok mennyisége mellett azok különféle fajtái is lehetnek olyan tényezők, amelyek az értékelők szigorát és az értékelés megbízhatóságát befolyásolják. Olyan értékelők, akik kimaradtak a feladatok közös gyakorlásából (*community of practice*), nagyobb valószínűséggel értékeltek alacsonyabb megbízhatósággal. Lim (2011) az értékelői következetességet (*intra-rater reliability*) vizsgálta az íráskészség mérésénél, összefüggésben az idővel és a tapasztaltság mértékével, míg Yan (2014) vizsgálata az értékelők kölcsönhatását célozta meg.

Az értékelő orientációjától a hierarchikus értékelői modellig

E tanulmány szempontjából legfontosabb kutatási irányt a hierarchikus értékelő modellhez (HRM) vezető, illetve ahhoz kapcsolódó írásművek jelzik. Mint alább kitűnik, a saját adatok elemzése során az értékelői (témavezetői vagy bírálói) szerepeknek a nemzetközi kutatásokhoz hasonló összetett, hierarchikus rendje bontakozott ki. Az értékelői hatás összetettségéről való gondolkodás a szakirodalomban három, egymásból láncszerűen fejlődő elmélettel jellemezhető.

Az értékelői orientáció

Az értékelői hatás szempontjából figyelemre méltó Brown és munkatársai (2005), továbbá Ducasse és Brown (2009), akik az értékelői orientációt vizsgálták, tudniillik azokat az aspektusokat, performanciatényezőket, amelyekre az értékelők odafigyelnek. Hasonlóképp, Eckes (2008) kategóriákat állított fel aszerint, hogy az íráskészség-értékelők mire figyelnek oda (és mire nem). Cumming és munkatársai (2002. 67.) az értékelők döntéshozatali mechanizmusainak vizsgálata során feljegyezték, hogy „míg az ESL/EFL hátterű íráskészség értékelők figyelmének középpontjában összességében inkább a nyelvtudás állt, mint a retorika vagy a gondolati tartalom, addig az angol anyanyelvű

értékelők figyelme kiegyenlítettebben oszlott meg a felsorolt szempontok között", vagyis korábbi tapasztalataik, képzésük stb. hatottak arra, mire figyelnek oda.

A szignáldetekciós elmélet

Az értékelői figyelem megoszlásához és irányához (orientációjához) kapcsolódó gondolatok precíz megfogalmazása a szignáldetekciós elmélet (*Signal Detection Theory*, SDT). Lényege, hogy az észlelés egyrészt a jel(enség) erejétől függ, másrészt attól, hogy az észlelő, jelen esetben az értékelő mire figyel fel, mire nem, és milyen módon oszlik meg figyelme. DeCarlo (1998, 2005) cikkeiben visszatekint az SDT kezdeteire, megvilágítva azt, hogy az SDT a hierarchikus értékelő modell előképének tekinthető (Patz és mtsai, 2002; DeCarlo, 2005), vagyis az SDT hatott az alább tárgyalt hierarchikus értékelő modell (*Hierarchical Rater Model*, HRM) kialakulására (DeCarlo, 2005; Patz és mtsai, 2002).

A hierarchikus értékelői modell

DeCarlo és munkatársai (2011) arra a következtetésre jutottak, hogy az íráskészség értékelésében eltérő szintek azonosíthatók. Az értékelői pontértékek nem szükségképp közvetlen indikátorai a vizsgázói teljesítményeknek (ideértve a szakdolgozatokat is), mert az a témavezetők és bírálók pontozásán keresztül valósul meg, vagyis eltérés van az adatstruktúrában elfoglalt helyük tekintetében. A modell szerint az értékelők a feladaton mérhető teljesítményt értékelik és nem közvetlenül az értékelt személyek (vizsgázók, szakdolgozatok szerzői) teljesítményét. DeCarlo és munkatársai (2011. 333.) hozzáteszik még, hogy több értékelő bevonása nem növeli meg az értékelés precizitását, viszont több feladat elvégeztetése igen. A modell hasznos, mert rávilágít arra, hogy az értékelők sem tekinthetők a performancia monolitikus tényezőjének, irodalma pedig azért releváns, mert mind a mért kompetencia, mind a performancia tényezői összetettek, sokrétűek, tovább tagolhatók alkotó elemeikre.

Amint azt DeCarlo (2005. 53.) saját kiindulópontjaként bemutatja, a mérési-értékelési szakma meglehetősen megosztott a tréning és vizsgáztatás során megkövetelt értékelői egyetértés fontossága tekintetében, a megbízhatóság értékelésének módszereiben, illetve a mérés során alkalmazandó pszichometriai modell tekintetében. A korabeli megosztottságot érzékelteti Linacre és munkatársai (2003) cikke is, akik korábbi tapasztalataik alapján bírálják a minél nagyobb értékelői egyetértésre való törekvést, mellyel az oktatási rendszert azonosítják. Linacre pozitív tapasztalatokat egészségügyi területen szerzett, ahol az értékelő független szakértő, pl. a diagnózisát felállító orvos, míg az oktatási rendszerben az egyéni látásmód fel kell oldódjon a kiadott értékelési irányelveknek való megfelelésben: „Az értékelői munka modellezése problematikus" (Linacre és mtsai, 2003. 928.). Felteszi kérdést, hogy az értékelőnek a független szakértő (*locally-independent expert*) vagy a pontozógép szerepét szánjuk. DeCarlo (2005. 54.) viszont a szintézis megteremtésének akadályát abban látja, hogy a mérési modelleknek jellemzően nincs erős kapcsolatuk a lehetséges pszichológiai modellekkel, ehhez viszont több ismeret kell arról, mit gondolnak, látnak az értékelők az értékelés során.

Az érvényesítés elméleti háttere

Értékelési eredmények érvényessége tekintetében a kutatás olyan eljárást követ, amelyben először az összes válaszadat konzisztenciáját (megbízhatóságát), illetve az értékelői megbízhatóságot szokás vizsgálni (Brennan, 2006. 5.). Ide tartozik a megfelelő pontozás is, Csikosnál (2020. 77.) skálázási validitás, vagyis olyan megbízhatósági problémák kezelése, amelyeknek egyértelmű negatív hatása van a pontértékek validitására (érvényességére). A megbízhatóság vizsgálatát később követheti az eredmények érvényessége mértéke megállapítását célzó további munka, ha a mérési eszközrendszer (skálák, értékelők) megbízhatósága kielégítő. Összességében elmondható, hogy megbízhatósága a válaszadatoknak van, míg érvényessége az osztályzatoknak, eredményeknek, értékeléseknek és azok következményeinek kell legyen. Amennyiben a megbízhatóság nem kielégítő, az eredmények érvényességét sem érdemes tovább vizsgálni. A megbízhatóság és érvényesség kapcsolata aszimmetrikus: A mérés megbízhatósága elvileg teljesülhet megnyugtató érvényesség nélkül is, de eredmények, konklúziók nem címkézhetők érvényesnek gyenge konzisztenciájú, megbízhatóságú válaszadatok alapján.

Az érvényesítés tudományos irodalmában természetesen nem ez az egyetlen eljárás a mérés-értékelés szakterületén belül; megemlíthető például Weir modellje (*Socio-Cognitive Model*), Bachman és Palmer modelljei (1996, 2010), Kane (1992, 2012) modellje (*Argument-based Model*), például Kimnél (2011), vagy időben még korábban Messick modellje (1989, 1995). Bár az utóbbiak egyértelműen csoportot alkotnak, közülük is Messick modellje tűnik a legnagyobb hatásúnak, nem csak a nyelvpedagógiában: Ahogy Bachman és Palmer sokat hivatkoznak Messickre, úgy Kane is Messick gondolatait követi, kiegészítve azokat. A csoport gondolati rokonsága, korai gyökerei okán e modelleket, az irányzatot klasszikus vagy bevett megközelítésnek is nevezhetjük, Fulcher (2014) szerint episztemológiainak nevezett filozófiai alapokkal, mivel a tudás (kompetencia), vagy más hasonló pszichológiai attribútum egyszerű tapasztalással nem felismerhető. (Borsboom és mtsai [2004] például a saját, ontologikus megközelítésükkel állítják szembe az episztemológiai megközelítést.)

A kutatás kérdései

A kutatás legfontosabb kérdése a kettős értékelés minősítése, amely azonban nem lehet a sorrendben az első kérdés, mert az érvényesítés módszertana szerint először a mérés megbízhatóságát, konzisztenciáját kell megnyugtató módon tisztázni.

1. A megbízhatóság, konzisztencia mutatói megfelelők-e, és lehetővé teszik-e az érvényesség vizsgálatát?
2. Mennyiben éri el szakdolgozatok kettős értékelése a célját? Az értékelői eltérések mivel magyarázhatók?

A kutatás módszerei

A módszerfejezet leírásába beletartoznak az intézményi (tanárképzési program) keretek, ahol a kutatás végbement, a felhasznált adatok köre és a szakdolgozati skálák bemutatása, a párhuzamos értékelési eljárás bemutatása, a szakdolgozati konstruktum értékelői aspektusai is, az elemzési módszerek és annak szakaszai ismertetése mellett.

A kutatás keretei

A kutatás kereteit az ELTE Angol-Amerikai Intézete végzős angoltanár (osztatlan tanárképzés, OTAK) hallgatói szakdolgozatainak értékelése határozta meg. Ennek része volt az a kb. 5% olyan angoltanár szakos hallgató is, akik választott témavezetőjük okán az Angol Nyelvpedagógia és az Angol Alkalmazott Nyelvészet Tanszékeken kívül, más intézeti tanszéken (programban, pl. anglisztika) írták szakdolgozatukat. Kívül estek a kutatás keretein azok az angoltanár szakos hallgatók, akik szakpárosításuk másik szakja keretében készítették dolgozataikat.

Az adatgyűjtés keretei és a feldolgozott adatok

Az 1. táblázat a 2011 ősze és 2019 ősze közti 17 félév összesen 310 dolgozat 4340 válaszadata alapján végzett értékelését tanúsítja, míg Dávid és Piniel (2018) elemzése csak a táblázatban szürkével jelölt, 2011–2013 közötti kezdeti időszakra terjedt ki. Hozzá kell tenni, hogy míg a cikkben feltett kutatási kérdések megválaszolásához megfelelők a 2011–2019 között felvett adatok, addig a 2020–2025 között gyűjtött adatok az értékelési rendszer változatlan alkalmazása mellett már új kérdéseket is felvetnek, így pl. a Covid hatása, a dolgozatok elektronikus benyújtása és értékelése már egy másik kutatáshoz tartozik.

1. táblázat. A szakdolgozók félévek szerinti megoszlása

Egyetemi félévek	Szakdolgozatok száma
2011. ősz	8
2012. tavasz	7
2012. ősz	29
2013. tavasz	17
2013. ősz	17
2014. tavasz	12
2014. ősz	21
2015. tavasz	15
2015. ősz	22
2016. tavasz	16
2016. ősz	14
2017. tavasz	27
2017. ősz	27
2018. tavasz	20
2018. ősz	15
2019. tavasz	33
2019. ősz	10
Összesen	310

A szakdolgozati skálák

A szakdolgozati skálák analitikus skálák, mert hét egymást kiegészítő részskála alkotja őket: *Kutatás módszerei és eljárásai* (ReMePr [1]), *Formai követelmények* (FoR), *Eredmények értelmezése* (IntF), *Elméleti és tapasztalati alapok* (ThEB), *Az angol nyelvhasználat minősége* (QuEL), *Írás és argumentáció minősége* (QuWr) és *Önálló munka minősége* (Ind). A skálák mindegyike ötfokozatú, skálánként 0–4 ponttal. (Megtekintésre arányskálákról van szó, mert minden skála 0 pontról indul.) Minden skálaponthoz kritériumleírás tartozik. A skálák az alárendelt szempontokkal együtt megtalálhatók a https://delp.elte.hu/otak_theses webhelyen, illetve Dávid és Piniel (2018) B függelékeként is.

A kettős értékelésben érintett szakdolgozatok jellemzői

A hallgatók a szaktárgyi tanítási gyakorlat tapasztalatait felhasználva készítik el szakdolgozatukat, amiben – nagyon általánosan fogalmazva – leírják, hogy milyen kutatási kérdésekre keresték a választ, miért az adott nyelvpedagógiai vagy alkalmazott nyelvészeti témát és annak kérdéseit választották, milyen módszerekkel keresték a választ a kérdéseikre, milyen eredményeket kaptak, továbbá milyen tanulságokkal szolgált a kutató tanári tapasztalatszerzés.

Mivel tanár szakos hallgatókról van szó, a dolgozat jellemző módon osztálytermi, iskolai, de mindenképp a képzés szempontjából releváns saját kutatás elvégzését és megírását foglalja magában. Egy (vagy több) jól körülhatárolható kérdésre kell választ adni úgy, hogy a dolgozat nem több, mint 75-80 ezer leütés, +10% terjedelmi korláton belül (https://delp.elte.hu/otak_theses).

A kettős párhuzamos értékelés

Az értékelés a skálák (ELTE Angol-Amerikai Intézete *Rating scales for the Evaluation of OTAK / MA theses*) alapján adott pontokkal, vagyis kvantitatív eljárással veszi kezdetét. Ebben a szakaszban az értékelés oly módon kettős, hogy az értékelők nem egymás után értékelik a dolgozatot, hanem önállóan, azonos időszakban, egymás pontozását és esetleges kézzel írt megjegyzéseit nem látva végzik feladatukat. Ezt a fajta kettős értékelést kettős párhuzamos értékelésnek nevezem.

Az értékelési eljárás

A 2011 óta folytatott eljárás szerint minden szakdolgozat kettős értékelésre kerül. Az értékelők egyike a bíráló, aki azonban nem a képző intézménytől független, hanem attól a kapcsolattól, amely a szakdolgozót és témavezetőjét összeköti. A bíráló a dolgozat témájában jártas oktató. A másik értékelő a témavezető, akinek bevonására erős igény mutatkozott már a skálafejlesztés során. A bíráló és a témavezető egymástól függetlenül tesznek javaslatot az érdemjegyre a pontkonverziós táblázat alapján.

Az értékelés következő szakaszában a koordinátor megkeresi azokat a dolgozatokat, ahol a javasolt érdemjegyek eltérést mutatnak. Egy jegy eltérés esetén felkéri a két

1 A mozaikszavak a szempontok angol nyelvű elnevezéseiből származnak, és segítenek értelmezni a 4. táblázat sorait.

értékelőt, egyezzenek meg a végső dolgozatjegyben, hiszen a tanulmányi rendszer egyetlen jegyet enged. Ha két jegy a különbség, a szakdolgozó a köztes jegyet kapja. Harmadik értékelőt csak akkor kell felkérni, ha a különbség két jegynél nagyobb, vagy az érdemjegyek közül az egyik elégtelen. Az egyszerű döntési szabályok megakadályozzák, hogy az egyeztetés vállalhatatlanul gyakori legyen. Ismételt pontozást nem végzünk, a kettős értékelés az érdemjegy egyeztetésével lezárul.

A szakdolgozati konstruktum értékelői aspektusai

A szakdolgozat konstruktuma a skálafejlesztéssel alakult ki. Részét képezik azok a feltételezett pszichológiai folyamatok, amelyekben, eltérő helyzetükből adódóan, a témavezető és a bíráló részt vesznek. A témavezető ismeri a hallgatót, jellemzően több konzultáción fogadta, részt vett a téma kiválasztásában és a kutatási kérdések megfogalmazásában, továbbá jellemzően olvasta a szakdolgozat piszkozati változatait. A bíráló viszont lehet, hogy nem is ismeri a dolgozat szerzőjét.

E különbségek fontosak, mivel a témavezető háttérismeretek birtokában jobban érti a szakdolgozó gondolatmenetét még akkor is, ha a szakdolgozat minősége elmarad attól, ami elvárható. A témavezető tud a szakdolgozó olyan elvégzett, illetve el nem végzett, a szakdolgozattal kapcsolatos feladatáról, esetleg attitűdbeli, szorgalmi aspektusokról, a dolgozatírás körülményeiről is, amelyek nem jelennek meg a szakdolgozatban, vagy nincsenek elég jól leírva, míg a bíráló nem támaszkodhat másra, mint ami a dolgozatban áll. Ennek következtében a témavezető nem mint független értékelő értékeli a dolgozatot, míg a bíráló nem is tehet mást, minthogy a leírt szövegből indul ki, és produktumként értékeli. Jól érzékelhető különbség van így a témavezető és a bíráló szellemi és érzelmi folyamatai között. E különbségek magyarázhatják a kutatás során lentebb azonosított különbségeket. A szakdolgozat konstruktuma tehát két megközelítésmód kompromisszuma. Bár a témavezető feladata is az, hogy a dolgozatot értékelje, továbbá az alkotás folyamatát értékelheti a szakdolgozati konzultáció osztályozásában, feltételezhető, hogy értékelésére hatnak a tapasztalatok, amiket a megíratás és konzultációk során szerzett.

A kettős értékelés itt tárgyalt megoldása megállja a helyét nemzetközi gyakorlatban. Példa lehet McQuade és munkatársai (2020), ahol a témavezető és bíráló súlya a kettős értékelésben azonos, és a megfigyelések is nagyon hasonlóak, vagy Alderson és munkatársai (1995), akik iránymutatásukat a brit vizsgaközpontok gyakorlatának kérdőíves felmérésére alapozták. A kettős értékelés megállja a helyét országos összehasonlításban is, hozzátéve, hogy a magyar nyelvű szakirodalom nem bővelkedik az értékelési skálák alkalmazása témájú vagy a kettős értékelési eljárás kutatásában (Kiszely, 2012, 2019).

A hazai értékelési eljárások rövid összefoglalása és egyben kitekintés lehet, hogy az angoltanár-képzést az OTAK program szerint folytató hat legjelentősebb egyetem gyakorlata jelentős eltéréseket mutat, amelyben a kettős értékelés tekintetében az itt tárgyalt kettős értékelés nem képviseli a többséget, de nem is az egyetlen. A vonatkozó honlapok és a szerző tájékozódása alapján van olyan eljárás, amelyben nincs kettős értékelés; van, ahol a kettős értékelést nem követi egyeztetés; van, ahol alkalmaznak explicit értékelési skálákat (deskriptorokkal), és van, ahol az oktatási rendszerben általános, 1–5-ig terjedő érdemjegy-skálát alkalmazzák csak, deskriptorok nélkül. Az ELTE-n belül a kettős értékelést folytató programok száma alacsony, viszont az elbírált angoltanári szakdolgozatok száma magas (ELTE BTK Tanulmányi hivatal, személyes közlés).

Az adatelemzés módszerei

Az elemzések első szakasza a nyerspontértékek konzisztenciájának (megbízhatóságának) vizsgálatát célozta a Chronbach-alfa kiszámításával (SPSS 27.0, IBM Corp., 2020). Ezt követte egyrészt a szakdolgozati skálák megfelelése (beválása) mértékének, másrészt a kettős értékelési eljárás érvényességének a vizsgálata.

A többdimenziós Rasch-elemzés (MFRM)

Az elemzések sorában fontos szerepet kapott a többdimenziós Rasch-elemzés a Facets szoftver segítségével (Linacre, 2014). A Facets a probabilisztikus (valószínűségi) módszercsalád egy paraméteres, Raschnak is nevezett ágához tartozik. Linacre (1989) fejlesztése lehetővé tette, hogy ne csak a szakdolgozatokat lehessen kalibrálni, hanem a mérés több más dimenzióját (változóját) is, esetünkben a skálákat és az értékelőket, illetve az értékelői szerepeket is. A dimenziókat a szociológia területéről származó *facet theory* alapján (Guttman és Greenbaum, 1998) Linacre faceteknek nevezte – innen a többdimenziós módszertan elnevezése is: *Many-facet Rasch Measurement* (MFRM). A probabilisztikus mérésre jellemző, logitban megadott mutatók elemzése a skálák megfelelése kérdésre adott választ: A skálák mennyire illeszkednek a probabilisztikus mérési modellhez?

A nyerspontértékek feldolgozása során a Facets az arányskálákat ordinális (rang-) skálákká alakítja. Nem feltételezi, hogy az egyes skálapontok azonos távolságra vannak egymástól, majd pedig a betáplált tényezők során intervallum típusú logit skálává alakítja őket át. A skálák megfelelését a logit alapú illeszkedési statisztikák alapján vizsgáltam, ezen belül az *infit* átlagos négyzetes (súlyozott) illeszkedési mutató (*infit meansquare*/*MNSQ*), illetve ennek standardizált variánsa, az *infit Z* érték alapján, valamint az outfit átlagos négyzetes (súlyozatlan) illeszkedési mutató (*outfit meansquare/MNSQ*), illetve ennek standardizált variánsa (*outfit Z*-standardizált érték) alapján. A *point-biserial* típusú korrelációt a Linacre (2017) által javasolt *Point-Measure* változat alapján vizsgáltam. A releváns elemzések alább olvashatók (3. táblázat).

A Rasch-módszertan minden dimenzióra meghatározza a küszöbértékeket, amely értékek felett a megmagyarázatlan szórás már nem elfogadható. Linacre ajánlását egy kicsit módosítva alkalmaztam. Feltételeztem, hogy az illeszkedési értékek normál eloszlásúak, ezért a skála, skálapont illeszkedési mutatója az átlag felett 2 szórásegységen (*standard deviation*, SD) túl már extrém, elfogadhatatlan értéknek tekinthető.

A vizsgálat szakaszai

Figyelembe véve a kutatási kérdések sorrendjét és az érvényesség vizsgálatának követelményeit, az elemzések a megbízhatósággal kapcsolatos (első) kutatási kérdés megválaszolásával kezdődtek, az alábbi szakaszok szerint:

1. szakasz: a vizsgált pontértékek megbízhatósága (konzisztencája).
2. szakasz: a témavezető és a bíráló összesített pontszámainak korreláltatása.
3. szakasz: a skálák megfelelésének vizsgálata.

Az elemzések a kettős értékeléssel, az értékelők egyetértésével, összességében az érvényesítéssel kapcsolatos második kutatási kérdés megválaszolásával folytatódtak.

4. szakasz: a témavezető és bíráló az összesített pontszám korrelációja.

5. szakasz: a pontszám és a javasolt érdemjegy korrelációja mind a témavezető, mind a bíráló esetében.
6. szakasz: a Fleiss-Kappa egyetértési mutató kiszámítása.
7. szakasz: a témavezetői és bírálói pontátlagok összevetése.

A korrelációkat kiegészítették páros t-próbák az átlageltérések vizsgálatára. Az összes elemzési szakasz (1–7) áttekinthetősége szempontjából hasznos azok táblázatos ábrázolása (2. táblázat).

2. táblázat. Az elvégzett konzisztencia- és korrelációs vizsgálatok táblázatos áttekintése

Elemzési szakasz	Kutatási kérdés	Eszköz	Az összevetés és elemzés alapja	
1	megbízhatóság kérdése	konzisztencia (SPSS)	az összes (310 x 2 x 7) pontérték konzisztenciája a mérésben	
2		korrelációk (SPSS)	témavezető összesített pontjai és javasolt jegye	bíráló összesített pontjai és javasolt jegye
3		többdimenziós Rasch (Facets)	az összes (310 x 2 x 7) pontérték a mérésben	
4	érvényesítés kérdése	korreláció (SPSS)	témavezető összesített pontjai	bíráló összesített pontjai
5		korrelációk (SPSS)	témavezető összesített pontjai és a végső érdemjegy	bíráló összesített pontjai és a végső érdemjegy
6		egyetértés (Fleiss-Kappa)	az összes értékelő pontjai	
7		páros mintás t-próba (SPSS)	az összes értékelő pontjai	

A kutatás eredményei

A kutatás első, megbízhatóságot és konzisztenciát érintő kérdésére mindenképp pozitív válasz adható. A Chronbach-alfa összesen 310 dolgozat adatainak betáplálásával elérte a 0,914 értéket, dolgozatonként 14 pontértékkel lehetett számolni, mivel az elemzés mindkét értékelő 7-7 pontjára kiterjedt. A bíráló pontjainak és a pontkonverzió szerinti javasolt érdemjegyek korrelációja r = 0,901 (N = 309), amitől kicsit elmarad a témavezető pontjainak és a konverziós táblázat szerinti érdemjegyeinek korrelációja r = 0,883 (N = 304), vagyis a konzisztenciát a pontok érdemjegyekre váltása fenntartja és tovább vetíti, megerősítve ez által a megbízhatóságot. Fontos eredményről van szó, mert az értékelők összesített pontjainak adott tartománya kell korreláljon a javasolt 1–5-ig érdemjeggyel.

A skálák megfelelése, beválása

Az értékelésben használt skálák megfelelése elsősorban a megbízhatóságról nyújt információt, de ha megfelelő, közvetve az érvényesség megítélését segíti. Általánosságban megállapítható (3. táblázat), hogy a skálák beváltották a hozzájuk fűzött mérési elvárásokat. Az illeszkedési statisztikák fényében a legtöbb skála megfelelő. Az infit (súlyozott) és outfit (súlyozatlan) mutatók, mind a meansquare változat, mind pedig a standardizált változat nem lépte túl a 2 szórásegység (SD) limitjét. A 3. táblázat a skálák megfeleléséről további tájékoztatást nyújt a megfigyelések száma (gyakoriság), a nyerspontok átlaga 0–4 pont között, a logit nehézségi érték, valamint a *Point-Measure* korreláció megfigyelt és elvárt értékéről is.

A skálák megfelelése tekintetében kivétel a *Formai követelmények* (FoR) skála, amely csak marginálisan fogadható el (3. táblázat). E skála *infit meansquare* értéke ugyan még épp megfelelő (1,23), viszont az *outfit* érték (1.29) már nem, mert túlmegy a 2 szórásegységen (SD), és már extrémnek tekinthető. Mivel az outfit súlyozatlan érték, érzékeny a nem várt, kiugró (*outlier*) válaszokra. Ha az *infit* és *outfit* standardizált (Z) értékeket vetjük össze, azok már nem jeleznek túl sok megmagyarázatlan szórást, már nem lépik túl a határértékeket. További információ a *Formai követelmények* skála egyes sávjainak megfelelőségéről a Függelékben található, mivel a 3. táblázat csak a skála egészéről nyújt információt.

3. táblázat. Az ELTE-Angol-Amerikai Intézet angoltanári skálák megfelelésének összefoglaló táblázata

Skálák	Gyakoriság	Nyers átlag	Nehézség	Hibaérték	InfitMS	InfitZ	OutfitMS	OutfitZ	Diszkrim.	PtMeas	PtElv.
QuWr	609	3,32	0,79	0,07	0,86	-2,54	0,83	-2,81	1,17	0,72	0,67
IntF	609	3,33	0,61	0,07	0,96	-0,76	0,93	-1,07	1,06	0,68	0,67
ReMePr	609	3,18	0,32	0,07	0,95	-0,78	0,92	-1,2	1,05	0,72	0,71
QuEL	609	3,1	-0,17	0,07	1,12	2,02	1,12	2,07	0,86	0,62	0,66
ThEB	609	3,35	-0,42	0,08	0,91	-1,58	0,93	-1,04	1,09	0,69	0,66
FoR	609	3,32	-0,48	0,07	**1,23**	3,61	**1,29**	**3,92**	**0,72**	0,60	0,68
Ind	609	3,55	-0,66	0,08	0,9	-1,36	0,87	-1,25	1,06	0,67	0,65
Átlagértékek		3,31	0,00	0,07	0,99	-0,20	0,98	-0,20	1,00	0,67	0,67
Szórás		0,13	0,53	0,00	0,12	2,00	0,15	2,20	0,12	0,04	0,01

A megbízhatóságot erősítő további tényezők

A skálák változtatás nélküli alkalmazása, továbbá a visszatérő értékelők pontozása mintegy összekapcsolta az adatcsomagokat, megteremtve azok összekötöttségét (*subset connection*), és lehetővé téve összevethetőségüket, amint azt a Facets szoftverbe épített ellenőrző algoritmus (Linacre 2014, 2017) megnyugtató módon jelezte is. Az adatok összekötöttségét lehetővé teszi és ezen keresztül a megbízhatóságot erősíti a kutatás közel teljes mértékben keresztezett szerkezete is: A témavezetők 82%-a (31/38 fő) egyben bíráló is volt, míg a bírálók 87%-a (27/31) egyúttal témavezetők is, vagy – megfordítva a fenti megállapításokat – az értékelők egy kisebb része vagy csak témavezető volt (18%), vagy csak bíráló (13%).

További összefüggések és az átlagok vizsgálata az érvényesség tekintetében

A megbízhatóságot és konzisztenciát érintő első kutatási kérdés megnyugtató tisztázása elvezet a második kutatási kérdéshez (érvényesség) kapcsolódó információ mérlegre tételéhez. Ennek egyik eleme két további korrelációs eredmény, majd az értékelői egyetértés vizsgálata, ezen belül a pontátlagok összehasonlítása volt.

A témavezetői és bírálói összpontszámok egymás közötti Pearson-korrelációja közepesnek minősíthető és statisztikailag szignifikáns ($r = 0{,}591$, $p < 0{,}01$) volt. Fontos továbbá, hogy a témavezető ($r = 0{,}681$, $p < 0{,}01$) és a bíráló pontszáma is ($r = 0{,}758$, $p < 0{,}01$) ugyanazzal a végső egyeztetett érdemjeggyel korrelál – nem a javasolt jegygyel – mintegy jelezvén előre, hogy a pontértékek különbségeinek vizsgálata nem megkerülhető. Mindez már arra is utal, hogy van különbség a témavezetők és bírálók pontozása között.

Az egyetértés mértéke az értékelők nagy száma miatt a Fleiss-Kappával volt jellemezhető, melynek értéke 0,181, meglehetősen alacsony, de szignifikáns ($p < 0{,}01$) volt, vagyis a mérés pontos volt, kis konfidenciaintervallummal. A magas belső konzisztencia mellett jelentkező különbségekre a témavezetői és bírálói összpontszám-átlagok és javasolt jegyátlagok mutatnak rá. Ezek az átlagok azt mutatják, hogy a bírálók szigorúbbak, a témavezetők pedig elnézőbbek az értékelői átlaghoz képest. A 4. táblázat különbségei, mind a pontátlagokat tekintve ($t = 6{,}368$, $df = 302$, $p < 0{,}001$), mind pedig a pontkonverziós táblázat szerint javasolt (nem egyeztetett) jegyek tekintetében ($t = 4{,}822$, $df = 302$, $p < 0{,}001$), szignifikánsak a plafoneffektus (*ceiling effect*) lehetséges kedvezőtlen hatása ellenére. Ez az eltérés önmagában nem is lenne meglepő, ha a bírálók és témavezetők teljesen külön csoport lennének. Az eltérések azonban mégis azért különösen érdekesek, mert túlnyomórészt ugyanazok a kollégák pontoznak, vagy témavezetőként, vagy bírálóként, jellemzően ugyanabban a félévben, vagyis a kutatás keresztezett szerkezetű. Valószínűtlen tehát, hogy a témavezetők magasabb pontátlagát csak eme csoport elnéző pontozása magyarázná. Mindez arra utal, hogy a skálák és értékelők dimenziói (facetjei) mellett tetten érhető a szórás egy további dimenziója is, ami az értékelői szerepek dimenziója lehet.

Az egyetértés mértéke az értékelők nagy száma miatt a Fleiss-Kappával volt jellemezhető, melynek értéke 0,181, meglehetősen alacsony, de szignifikáns ($p < 0{,}01$) volt, vagyis a mérés pontos volt, kis konfidenciaintervallummal. A magas belső konzisztencia mellett jelentkező különbségekre a témavezetői és bírálói összpontszám-átlagok és javasolt jegyátlagok mutatnak rá. Ezek az átlagok azt mutatják, hogy a bírálók szigorúbbak, a témavezetők pedig elnézőbbek az értékelői átlaghoz képest. A 4. táblázat különbségei, mind a pontátlagok egyszerű összegeit tekintve ($t = 6{,}368$, $df = 302$, $p < 0{,}001$), mind pedig a pontkonverziós táblázat szerint javasolt (nem egyeztetett jegyek tekintetében ($t = 4{,}822$, $df = 302$, $p < 0{,}001$), szignifikánsak a plafoneffektus (ceiling effect) lehetséges kedvezőtlen hatása ellenére.

4. táblázat. Releváns átlagok összevetése

Osztályozási szempontok		Átlagértékek	N	Szórás	Átlag standard hibája
Összpontszám alapján	témavezetők	23,72	303	4,390	0,252
	bírálók	22,17	303	4,926	0,283
Jegyek alapján	témavezetők	4.64	303	1,247	0,072
	bírálók	4,30	303	0,787	0,045

Diszkusszió és összefoglalás

A skálák megbízhatósága egyrészt azt jelentette, hogy a skálák megállták a helyüket a szakdolgozatok értékelésében, másrészt azt, hogy a ponteredmények alapján meghozott döntések érvényességének további vizsgálata lehet a kutatás folytatása. Bár a kettős értékelés – mint fentebb láthattuk – logisztikai szempontból jelentős teher, a kettős értékelésre, a dolgozatonkénti 2 x 7 (14) pontértékére szükség van, mert csupán egy értékelő alkalmazása esetén a Chronbach-alfa (ha csak a bíráló értékelne a 7 szempont alapján) 0,777 lenne, míg 0,775, ha csak a témavezető tenné ugyanezt. A kettős értékelésre tehát szükség van, mert erősíti a mérés konzisztenciáját. Megállapítható az is, hogy a skálák többségének megfelelősége jó, ami megerősíti az értékelési rendszer megbízhatóságáról alkotott képünket, míg a *Formai követelmények* skála nem egyértelmű megfelelése, illetve az egyes sávok problémái (ld. a Függelékben) valamelyest gyengíti azt.

További összegzés – és egyben kitekintés –, hogy a bíráló és témavezető közötti egyetértés (vagy egyet nem értés) Fleiss-Kappával mért alacsony értéke és a bírálói és témavezetői kör igen jelentős átfedése között látható ellentmondás feszül – mintha az értékelők önmagukkal nem értenének egyet. Sokkal hihetőbb az az értelmezés, miszerint a kettős párhuzamos értékelés működőképessége mellett a témavezetők és bírálók megközelítésmódja lehet eltérő, ami segít feloldani az ellentmondást. Az átlagpontszámaik között megfigyelt szignifikáns különbség (4. táblázat) arra utal, hogy a különbség az értékelői szerepüktől függ, vagyis hogy adott dolgozat kapcsán éppen témavezetésről vagy bírálói szerepről van szó. Nagy a valószínűsége, hogy e két értékelői csoport eltérő szemléletmódja, szellemi folyamatai és a szakdolgozóhoz való viszonya magyarázzák a pontkülönbségeket, a vizsgáztatásban tipikusnak mondható szigorúság (*severity*) vagy elnézőség (*leniency*) különbségei mellett. A fenti elemzések eredményei az értékelő dimenziónak hierarchikusan alárendelt *szerep* dimenzió létezésre utalnak, emlékeztetve az irodalmi áttekintésben tárgyalt hierarchikus érékelői modellre (DeCarlo, 1998, 2005; DeCarlo és mtsai, 2011).

Az ellentmondás a szakdolgozati értékelői konstruktum figyelembevételével feloldható. Mivel a témavezető és bíráló feladata, nézőpontja a maguk értelmezésében eltér, nem reális elvárás az idegennyelvtudás-mérés terén elvártakhoz hasonló egyetértés, ahol az egymás mellé rendelt értékelők feladata azonos, a mérni kívánt konstruktum is azonos, ugyanazokat a nyelvi jellemzőket kell figyeljék (értékelői orientáció) és ugyanazokat a skálákat kell pontozzák. Ha viszont a két értékelő szerepe eltér, értékelői konstruktumaik is eltérnek és nem lehet arra számítani, hogy az értékelői egyetértés nagy mértékű lesz. Hasonlóképp, amint pl. egyes nyelvvizsgák esetében is feltételezhető, ahol az egyik értékelő holisztikusan értékel, míg a másik analitikus skálák alapján pontoz, nem teljesülhet a magas szintű értékelői egyetértés.

A kutatás szakmai értéke

Ez a cikk annyiban járulhat hozzá ismereteinkhez, hogy egyrészt bizonyítja, a kettős párhuzamos értékelés, ha jelentős erőforrásokat is köt le, megbízható módon folytatható tartósan, másrészt megmutatja, hogy feltételezhetően az értékelői dimenzión (faceten) belül, illetve a mellett léteznie kell az értékelési helyzettől függő értékelői szerep dimenziónak is. Ha az értékelői konstruktum engedi, megjelenik, ha a konstruktum nem engedi meg, nem érdemes számolni vele – jelenléte akkor problematikus, ha a konstruktum ugyan nem engedi meg, mégis megjelenik. A kutatás értékét az adja, hogy a szerep dimenzió az értékelői dimenzióhoz kapcsolható szórás egy része, eleme, melyet akkor érdemes külön számítani, ha nem tekinthető a mérendő konstruktum részének.

A kutatás korlátai és a gyakorlati lehetőségek

Mivel a korrelációs értékek nem értelmezhetők ok-okozati viszonyként, hanem csak adott konstellációként, kvalitatív eszközökkel végzett kutatás igazolhatja, hogy a kollégák mint témavezetők tényleg azért elnézőbbek, és azért adnak jobb jegyeket, mert az értékelői konstruktum megengedi, és hogy tényleg nem csak azt osztályozzák, ami épp le van írva a dolgozatban, vagy pedig pedagógiai motiváció okán teszik ugyanazt. Ugyanakkor az is elképzelhető, hogy nem nyílnának meg a motivációjukat firtató kutató előtt.

Ha a *Formai követelmények* skála illeszkedése nem javul, a mögöttes gondolkodás elsősorban figyelmet, további tréninget és kutatást igényel, melyben kvalitatív módszerekkel választ kaphatunk arra a kérdésre, mi okozza a skála itt megfigyelt illeszkedési problémáját. Ha a tréning nem vezetne megfelelő eredményre, a skála elemzésekre épülő módosítása is a lehetséges forgatókönyvek egyike, ez esetben azonban a többi jól megfelelt skálát változatlan formában mint horgonyt (*anchor*) kell alkalmazni, és az új szövegezésű skálát moderálni és pretesztelni kell. Bár Dávid és Piniel (2018) korábbi elemzésükben az illeszkedési problémát már azonosították, a skála mutatói javuló tendenciát mutatnak az újabb, itt publikált elemzések fényében: Az illeszkedési mutatók már marginálisan elfogadhatók. A javuló mutatók oka lehet, hogy az értékelők utóbb jobban alkalmazzák a skálákat; így az áttervezés, a deskriptorok módosítása tehát nem abszolút prioritás.

Köszönetnyilvánítás

Köszönettel tartozom minden kollégámnak az évek hosszú során elvégzett értékeléseikért, akik így hozzájárultak a szakdolgozatok minőségbiztosítási rendszere működéséhez, fenntartásához, s így lehetővé téve e tanulmány megírását is.

Irodalom

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

Bachman, L. F. & Palmer, A. S. (1996). *Language Testing in Practice.* Oxford University Press.

Bachman, L. F. & Palmer, A. S. (2010). *Language Assessment in Practice.* Oxford University Press.

Bond, T. G. & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Lawrence Erlbaum Associates.

Borsboom, D., Mellenbergh, G. J. & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061–1071. DOI: 10.1037/0033-295X.111.4.1061

Bramley, T. & Dhawan, V. (2011). Investigating and reporting information about marker reliability in high-stakes external school examinations. Konferencia-összefoglaló. *European Conference on Educational Research, Berlin*. https://www.cambridgeassessment.org.uk/Images/111868-investigating-and-reporting-information-about-marker-reliability-in-high-stakes-external-school-examinations-.pdf

Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement. In Brennan, R. L. (szerk.), *Educational measurement.* 4. kiadás. American Council on Education/Praeger. 1–16.

Brooks, V. (2004). Double marking revisited. *British Journal of Educational Studies, 52*(1), 29–46. DOI: 10.1111/j.1467-8527.2004.00253.x

Brown, A. (2003). *Legibility and the rating of second language writing. IELTS reports 4.* International English Language Testing System (IELTS). 131–151. https://ielts.org/researchers/our-research/research-reports/legibility-and-the-rating-of-second-language-writing-an-investigation-of-the-rating-of-handwritten-and-word-processed-ielts-task-2-essays

Brown, A., Iwashita, N. & McNamara, T. (2005). An examination of rater orientations and test-taker performance on English-for-academic-purposes speaking tasks. *ETS Research Report Series,* (1), i–157. DOI: 10.1002/j.2333-8504.2005.tb01982.x

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Cambridge University Press.

Cumming, A., Kantor, R. & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal, 86*(1), 67–96. DOI: 10.1111/1540-4781.00137

Csíkos, Cs. (2020). *A neveléstudomány kutatásmódszertanának alapjai.* ELTE Eötvös Kiadó.

Dávid, G. & Piniel, K. (2018). Establishing categories in the design of rating scales for MA in-ELT theses. *Working Papers In Language Pedagogy, 12*. 55–82. https://doi.org/10.61425/wplp.2018.12.55.82

DeCarlo, L. T. (1998). Signal Detection Theory and Generalized Linear Models. *Psychological Methods, 3*(2), 186–205. DOI: 10.1037/1082-989X.3.2.186

DeCarlo, L. T. (2005). A model of rater behavior in essay grading based on signal detection theory. *Journal of Educational Measurement, 42*(1), 53–76. DOI: 10.1111/j.0022-0655.2005.00004.x

DeCarlo, L. T., Kim, Y. & Johnson, M. S. (2011). A hierarchical rater model for constructed responses with a signal detection rater model. *Journal of Educational Measurement, 48*(3), 333–356. DOI: 10.1111/j.1745-3984.2011.00143.x

Ducasse, A. M. & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing, 26*(3), 423–443. DOI: 10.1177/0265532209104669

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*(2), 155–185. DOI: 10.1177/0265532207086780

Eötvös Loránd University (é. n.). OTAK theses. https://delp.elte.hu/otak_theses

Fulcher, G. (2013). Philosophy and Language Testing. In Kunnan, A. J. (szerk.), *The Companion to Language Assessment,* 1431–1451. Wiley–Blackwell. DOI: 10.1002/9781118411360.wbcla032

Guttman, R. & Greenbaum, C. W. (1998). Facet theory: Its development and current status. *European Psychologist, 3*(1), 13–36. DOI: 10.1027/1016-9040.3.1.13

Hallinger, P. (2011). A review of three decades of doctoral studies using the principal instructional management rating scale. *Educational Administration Quarterly, 47*(2), 271–306. DOI: 10.1177/0013161X10383412

IBM Corp. (2020). *IBM SPSS Statistics for Windows* (Version 27.0) [Számítógépes szoftver].

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527–535. DOI: 10.1037/0033-2909.112.3.527

Kane, M. (2012) Validating Score Interpretations and Uses: Messick Lecture, The Language Testing Research Colloquium, Cambridge, April 2010. *Language testing, 29*(1), 3–17. DOI: 10.1177/0265532211417210

Kim, H. J. (2015). A qualitative analysis of rater behavior on an L2 speaking assessment. *Language Assessment Quarterly, 12*(3), 239–261. DOI: 10.1080/15434303.2015.1049353

Kim, Y-H. (2010). An argument-based validity inquiry into the empirically derived descriptor-based diagnostic (EDD) assessment in ESL academic writing. *Doktori értekezés.* University of Toronto. https://hdl.handle.net/1807/24786

Kiszely, Z. (2012). Egy anglisztika szakon használt szakdolgozati értékelési skála megújításának alapelvei és használatának első eredményei. In Sárdi, C. (szerk.), *A felsőoktatás-pedagógia kihívásai a 21. században*. Eötvös József Könyvkiadó.

Kiszely, Z. (2019). Értékelési skálák használata a beszédkészség és íráskészség vizsgákon. *Modern Nyelvoktatás, 25*(3–4), 120–135. https://ojs.elte.hu/modernnyelvoktatás/article/view/1480

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language testing, 19*(1), 3–31. DOI: 10.1191/0265532202lt218oa

Lamprianou, I., Tsagari, D. & Kyriakou, N. (2023). Experienced but detached from reality: Theorizing and operationalizing the relationship between experience and rater effects. *Assessing Writing, 56*(100713), 1–14. DOI: 10.1016/j.asw.2023.100713

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing*, *28*(4), 543-560. https://doi.org/10.1177/0265532211406422

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Mesa Press.

Linacre, J. M. (2014). *FACETS* (Version 3.71.4) [Számítógépes szoftver]. https://www.winsteps.com/facets.htm

Linacre, J. M. (2017). *A user's guide to FACETS Rasch-model computer programs* (Version 3.80). https://www.winsteps.com/facets.htm

Linacre, J. M., Patz, R. J. & Donoghue, J. R. (2003). The Hierarchical Rater Model HRM from a Rasch perspective. *Rasch Measurement Transactions, 17*(2), 928. https://www.rasch.org/rmt/rmt172k.htm

Lukácsi, Z. (2021). Developing a level-specific checklist for assessing EFL writing. *Language Testing, 38*(1), 86–105. DOI: 10.1177/0265532220916703

Luoma, S. (2004). *Assessing speaking*. Cambridge University Press. DOI: 10.1017/CBO9780511733017

Martens, F. L. (1979). A scale for measuring attitude toward physical education in the elementary school. *The Journal of Experimental Education, 47*(3), 239–247. DOI: 10.1080/00220973.1979.11011688

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

McQuade, R., Kometa, S., Brown, J., Bevitt, D. & Hall, J. (2020). Research project assessments and supervisor marking: maintaining academic rigour through robust reconciliation processes. *Assessment & Evaluation in Higher Education, 45*(8), 1181–1191. DOI: 10.1080/02602938.2020.1726284

Meadows, M. & Billington, L. (2005). *A review of the literature on marking reliability*. National Assessment Agency. https://assets.publishing.service.gov.uk/media/5a820a57e5274a2e87dc0d5a/0505_Meadows_and_Billington_CERP_RP.pdf

Messick, S. (1989). Validity. In Linn, R. L. (szerk.), *Educational Measurement.* American Council on Education/Macmillan. 13–103.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist, 50*(9), 741–749. DOI: 10.1037/0003-066X.50.9.741

Ofqual (2014a). *Review of double marking research*. https://assets.publishing.service.gov.uk/media/5a82b3efed915d74e623734f/2014-02-14-review-of-double-marking-research.pdf

Ofqual (2014b). *Review of marking internationally*. https://webarchive.nationalarchives.gov.uk/ukgwa/20141031163546/http://ofqual.gov.uk/documents/review-of-marking-internationally

Partington, J. (1994). Double-marking students' work. *Assessment & Evaluation in Higher Education, 19*(1), 57–60. DOI: 10.1080/0260293940190106

Patz, R. J., Junker, B. W., Johnson, M. S. & Mariano, L. T. (2002). The hierarchical rater model for rated test items. *Journal of Educational and Behavioral Statistics, 27*(4), 341–384. DOI: 10.3102/10769986027004341

Pollitt, A. (1991). Response to Charles Alderson's paper: 'Bands and scores'. In Alderson, J. C. & North, B. (szerk.), *Language testing in the 1990s: The communicative legacy.* Macmillan. 87–94.

PTMIK (2002). *Közös Európai Referenciakeret: Nyelvtanulás, nyelvtanítás, értékelés*. Pedagógustovábbképzési Módszertani és Információs Központ.

Safari, F. & Ahmadi, A. (2023). Developing and evaluating an empirically based diagnostic checklist for assessing second language integrated writing. *Journal of Second Language Writing, 60*, Article 101007. DOI: 10.1016/j.jslw.2023.101007

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25*(4), 465–493. DOI: 10.1177/0265532208094273

Sideridis, G. & Padeliadu, S. (2013). Creating a brief rating scale for the assessment of learning disabilities. *Journal of Learning Disabilities, 46*(2), 115–132. DOI: 10.1177/0022219411407924

Steele, J. & Shaw, M. (2022). Exploring the value of double marking in dissertation assessments. [Preprint]. *EdArXiv*. DOI: 10.35542/osf.io/ug7yb

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press. DOI: 10.1017/CBO9780511732997

Weir, C. J. (2005). *Language Testing and Validation: An Evidence-based Approach*. Palgrave Macmillan. DOI: 10.1057/9780230514577

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing, 10*(3), 305–319. DOI: 10.1177/026553229301000306

Williams, L. & Kemp, S. (2019). Independent markers of master's theses show low levels of agreement. *Assessment & Evaluation in Higher Education, 44*(5), 764–771. DOI: 10.1080/02602938.2018.1535052

Winke, P., Gass, S. & Myford, C. (2012). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing, 30*(2), 231–252. DOI: 10.1177/0265532212456968

Yan, X. (2014). An examination of rater performance on a local oral English proficiency test: A mixed-methods approach. *Language testing, 31*(4), 501–527. DOI: 10.1177/0265532214536171

Yanosky II, D. J., Schwanenflugel, P. J. & Kamphaus, R. W. (2013). Psychometric properties of a proposed short form of the BASC teacher rating scale–preschool. *Journal of Psychoeducational Assessment, 31*(4), 351–362. DOI: 10.1177/0734282912456969

Függelék

A *Formai követelmények* skála egyéb tekintetben, a klasszikus tesztelmélettől vett mutatók szerint sem jeleskedik. Ez a skála egyben a leggyengébben diszkriminál (0,72) a többi között, vagyis a szakdolgozóknak adott magas/alacsony pontértékek nem tesznek számottevő különbséget a dolgozatok között. Megfigyelhető továbbá az is, hogy a Point-Measure korreláció mért értéke a legalacsonyabb az többi skála között (0,60), továbbá elvárt értéke magasabb (0,68), mint a mért érték. További fontos mutató is ellentmondásos: Ez a skála a második „legkönnyebb" skála, átlagos logit nehézségi értéke (-0,48) alapján, azonban a nyerspontátlagok tekintetében már közepesnek tekinthető (3,32), vagyis a Facets átrendezi a szempontok nehézségi nyerspont-átlag rangsorát, kimutatva az eltérő kalibrált nehézséget.

Mivel a 3. táblázat illeszkedési értékei skálánként összesített átlagos értékek, a táblázatból nem olvasható ki a skálák egyes sávjainak (skálapontjainak) illeszkedése; ily módon a *Formai követelmények* skála „belső" részleteiről sem itt lehet tájékozódni, hanem az 5. táblázatból. Innen azt tudjuk meg, hogy 2. és 3., illetve a 3. és 4. skálapontok logit értékei közötti átlagos távolság továbbra sem elég nagy a Bond és Fox (2001) javasolta értékhez (1,4) képest, de a Dávid és Piniel (2018) által közölt korábbi értékeknél mindenképp jobb. A skálapontok outfit értékei is magasak; az 1. és 3. skálapontok tekintetében az elfogadhatóság határán mozognak.

5. táblázat. A Formai követelmények skála sávjainak illeszkedése

Sávok	Logit átlagérték	Outfit átlagos négyzetes illeszkedési mutató
0	-3,17	0,4
1	0,22	1,4
2	1,83	1,3
3	3,11	1,4
4	3,99	1,2

A *Formai követelmények* skála (FoR) Dávid és Piniel (2018) korábbi elemzéseiben sem volt problémamentes, akkor csak 78 dolgozat alapján. A szerzők korábbi értelmezése most is igaz lehet: A kollégáknak kétségeik lehetnek a formai követelmények relevanciája, értéke tekintetében, amelyeket mindazonáltal meg kellett követeljenek, hiszen az egyetemi szakdolgozatot kutatási színvonalon és szakszerűen kell elkészíteni, a formai követelmények betartásával.

Absztrakt

A kutatás a kettős párhuzamos értékelés megvalósítását vizsgálja az angoltanári szakdolgozatok esetében az ELTE Angol-Amerikai Intézetében, 2011–2019 között. Az eszközök a klasszikus, nyerspontokra épülő megbízhatósági, korrelációs elemzés és pontátlagok összevetése volt, kiegészítve a többváltozós, probabilisztikus Rasch-módszerrel. A pontozás konzisztenciája, megbízhatósága megfelelt az elvárásoknak, megteremtve így a feltételeket az érvényesség vizsgálatához, azonban az értékelők egyetértési mutatója alacsony volt. A magas belső konzisztencia mellett jelentkező szignifikáns különbségekre a témavezetői és bírálói összpontszám-átlagok és korrelációk mutatnak rá, amely különbség nem is lenne meglepő, ha a témavezetők és bírálók teljesen külön csoport lennének. Az eltérések mégis érdekesek, mert túlnyomórészt ugyanazok a kollégák pontoznak, vagy témavezetőként, vagy bírálóként, és fordítva, jellemzően ugyanabban a félévben, amint azt a kutatás itt leírt szerkezete mutatja. Az egyetértés alacsony értéke és a témavezetői-bírálói körök jelentős átfedése közötti ellentmondás – mintha az értékelők önmagukkal nem értenének egyet – az értékelői szerepek magyarázatával oldható fel: A témavezető és bíráló eltérő konstruktumaik szerint, nem azonos szemlélettel értékeli a dolgozatokat. Szellemi folyamataik és a szakdolgozóhoz való viszonyuk különbségei magyarázzák a pontkülönbségeket a vizsgáztatásban tipikusnak mondható szigorúság (*severity*) vagy elnézőség (*leniency*) különbségei mellett. E tanulmány bemutatja, hogy az értékelői hatás komplex, hierarchikus elrendezésű, összetevői nem csupán a szigor vagy elnézőség tényezője, dimenziója lehet, hanem abban az eltérő értékelői feladatkörökből származó *értékelői szerep* tényező is közrejátszhat.

Kulcsszavak: kettős párhuzamos értékelés, szakdolgozatok értékelése, értékelői szerepek, írott performancia