MANYE
MAGYAR ALKALMAZOTT NYELVÉSZEK ÉS NYELVTANÁROK EGYESÜLETE
HUNGARIAN ASSOCIATION OF APPLIED LINGUISTS AND LANGUAGE TEACHERS

MODERN NYELVOKTATÁS

ISSN: 1219-638X https://ojs.elte.hu/modernnyelvok/

2026 | *Modern Nyelvoktatás, 32*(1), 54–68.
https://doi.org/10.51139/monye.2026.1.54.68

TANULMÁNY

# ANALITIKUS SKÁLÁK ALKALMAZÁSA AZ ANGOLTANÁRI SZAKDOLGOZATOK ÉRTÉKELÉSÉBEN

THE APPLICATION OF ANALYTICAL SCALES IN THE RATING OF MA THESES IN TEACHING ENGLISH AS A FOREIGN LANGUAGE (TEFL)

**Dávid Gergely András**
david.gergely@btk.elte.hu
Eötvös Loránd Tudományegyetem, Bölcsészettudományi kar, Angol-Amerikai Intézet,
Angol Nyelvpedagógia Tanszék
https://orcid.org/0000-0003-2280-5451

## KIVONAT

E tanulmány célja az értékelési skálák megfelelésének vizsgálata az angoltanári szakdolgozatok esetében az ELTE Angol-Amerikai Intézetben 2011 és 2019 között. Az eszköz a klasszikus, nyerspontokra épülő megbízhatósági, korrelációs elemzés és pontátlagok összevetése volt, kiegészítve a többváltozós probabilisztikus Rasch-módszerrel. Az eredmény a pontozás konzisztenciája tekintetében kiemelkedően jó volt. A pontkonverzió megőrzi az értékelés következetességét az érdemjegyekben, viszont az értékelők egyetértési mutatójának alacsony volta arra enged következtetni, hogy a témavezető és a bíráló nem azonos szemlélettel értékeli a dolgozatokat.

**KULCSSZAVAK:**
*Analitikus skálák, skálák megfelelése, értékelők, szakdolgozatok értékelése, írott performancia*

## ABSTRACT

The paper examines the performance of thesis scales as part of the evaluation of TEFL MA theses at the School of English and American Studies, Eötvös Loránd University, between 2011 and 2019. The instruments include classical analyses of raw ratings for consistency, correlations and significant mean differences between assessors and supervisors and Many-facet Rasch Measurement. The results show a very high level of consistency in responses, with score conversion preserving the consistency of scoring in the marks awarded, but the agreement between the assessors was low, suggesting a difference between the approach of assessors and supervisors.

**KEYWORDS:**
*Analytical rating scales, scale performance, raters, assessment of theses, written performance*

## BEVEZETÉS

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

A tanulmány szerzőjének hite szerint bármilyen értékelési rendszert időről-időre vizsgálni kell, mert az értékelés nyomon követésére gyakran nem jut elég idő, energia és figyelem. Tisztázni kell, hogy a mérés eszközei mennyiben váltják be a hozzájuk fűzhető, jellemző szakmai elvárásokat (megfelelés, beválás), míg a kapcsolódó értékelések tekintetében fontos mérlegre tenni az értékelés konzisztenciáját (megbízhatóságot). Esetünkben az értékeltetési eljárás vizsgálata egyúttal az angoltanári OTAK-programban folytatott kettős értékelés mérlegre tételét is jelentené, tekintettel arra, hogy annak megbízhatóságát, értelmét egyes források (pl. Meadows & Billington, 2005) a szükséges ráfordítások okán időről-időre vitatják. A kettős értékelés kérdésének megválaszolását azonban terjedelmi okokból e tanulmány legfeljebb csak előkészítheti.

A tanulmány további célja, hogy bemutassa az értékelés eredményét befolyásoló legfőbb performanciatényezőket és rámutasson arra, hogy az értékelői hatás maga sem homogén tényező, illetve specifikusan arra, hogy az értékelői hatás a szakdolgozati bírálók és témavezetők szerepei mentén eltér a kettős értékelés keretein belül. A tanulmány mérlegre teszi azt is, hogy szükséges-e változtatni a skálákon.

Dávid és Piniel (2018), a skálák fejlesztői korábbi kutatásának is fontos eleme volt már, hogy megvizsgálják, mennyire váltak be az akkor elkészült skálák, ebbe azonban csak az első 78 dolgozat kerülhetett be (ld. szürke kiemelés, 1. táblázat). Az itt vizsgált időszakban viszont több mint 300 dolgozat értékelése készült el, így tehát elvárható, hogy a skálák több éves alkalmazás után ismét megmérettessenek.

**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 |

## AZ ÉRINTETT SZAKDOLGOZATOK JELLEMZŐI

A hallgatók a szaktárgyi tanítási gyakorlat tapasztalatait felhasználva szakdolgozatot készítenek, amelyben – nagyon általánosan fogalmazva – megírják, 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 kutatótanári tapasztalatszerzés.

A tanárszakos hallgatók szakdolgozata jellemző módon osztálytermi, iskolai, de mindenképp a képzés szempontjából releváns saját kutatás elvégzését és megírását foglalja magában. Jellemző módon egy (vagy több) jól körülhatárolható kérdésre kell választ adni úgy, hogy a dolgozat belül maradjon 75–80 ezer leütés, illetve a 40 oldal + 10% terjedelmi korláton (https://delp.elte.hu/otak_theses). Az értékelés 2011 óta 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 kettős, önálló (párhuzamos) értékelés folyik.

## IRODALOM

A szakirodalmat az alábbi kérdéskörökre bontva, a terjedelmi korlát miatt szelektíven tárgyalom: a skálák megbízhatóságával kapcsolatos szkepticizmus, a kettős értékelés helyzete, valamint a két legfontosabb performanciahatás.

### 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 szerkezeti, tapasztalati 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, p. 90), vagy hét (Alderson et al., 1995, p. 111). Luoma (2004, p. 80) is rámutat, hogy a vizsgarendszerekben alkalmazott skálák sávjainak jellemző száma 4–6. Weigle (2002, p. 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 (Pedagógus-továbbképzési Módszertani és

Információs Központ, PTMIK, 2002, p. 239), míg Luoma (2004, p. 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, továbbá a nagyszámú sáv a legtöbb, amire az értékelő megfelelő módon figyelni tud, hogyan lehetséges 25 állítást nyomon követni (pl. Yanosky et al., 2013), ha minden állításhoz négyfokozatú skála kapcsolódik?

A megbízhatóság követelményéhez szorosan kapcsolódik a kettős értékelés kérdése. Mivel a legújabb szakirodalmi kritika szerint kérdéses, hogy a kettős értékelés mennyiben javítja a megbízhatóságot (Bramley & Dhawan, 2011; Steele & Shaw, 2022; Williams & Kemp, 2018), az ELTE angoltanári program szakdolgozatainak kettős értékelésének mérlegre tétele fontos elvégzendő feladat, azonban e tanulmányban a terjedelmi korlátok miatt erre a kérdésre nem lehet kitérni.

## A legfontosabb performanciatényezők

A szakdolgozatok esetében – miként az ún. szubjektív értékelésnél is – feltételezhető, hogy az eredményre számos performanciatényező hat. Ezek közül alább a skálák hatását és az értékelői hatást tekintem át, amely tényezőkről tapasztalati úton tudjuk, hogy közvetlen módon és jelentős mértékben befolyásolják az értékelést.

### *Skálák vagy listák az értékelésben?*

Dávid és Piniel (2018) szándéka az volt, hogy klasszikus (táblázatos) elrendezésű analitikus skálát tervezzen, melyben az oszlopok az értékelés szempontjainak, míg a sávok a vizsgált készség szintjeinek feleltethetők meg. Lukácsi (2021) azonban a KER-ből kiindulva kifejti, hogy az analitikus skálák helyett listák (checklist) alkalmazása reális alternatíva (PTMIK, 2002, pp. 233–234). Listák használata esetén az értékelő jellemzően a táblázatos elrendezésű deskriptorok (kritériumleírások) helyett egy listát tanulmányoz, ahol minden állítás a követelmények egy-egy eleme. 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. A KER megfogalmazása szerint – metaforikusan – ez „vízszintes" orientáció (p. 234): a lista tételei szintkövetelményként értelmezhetők, amelyeknek a teljesítmény vagy megfelel, vagy nem. Az ELTE-program kontextusában ez megfelel/nem felel meg (bináris) osztályzást jelentene.

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

*Listás, skálás és vegyes megoldások*

A listás megközelítés lehetséges vonzereje még, hogy a lista végigvezeti az értékelőt a szempontokon, ahol egyszerű eldöntendő kérdéseket kell megválaszolni. Ez a megközelítés egybecseng nyelvvizsgák és más idegen nyelvi szűrővizsgák legfőbb feladatával: bináris döntést kell hozniuk arról, hogy a vizsgázó (jelölt) megfelel-e adott követelményeknek vagy nem. Lukácsi (2021) a listás megoldás sikeres bevezetéséről számol be adott nyelvvizsgarendszer esetében, ahol a sikeresség feltétele a szintkövetelmények minimumának teljesítése. Sikeres példákat is hoz, pl. Kim (2010) és Kim (2011).

Az idegennyelv-tudás mérése szakterület mellett a listás megoldás a pszichológia, az egészségügy (fogyatékosság) (pl. Sideridis & Padeliadu, 2011), az oktatás-menedzsment (pl. Hallinger, 2010) terén is megtalálható. A szakirodalom áttekintése során azonban feltűnő volt, hogy kevés olyan további tisztán listás megoldást lehetett találni, melyekre Lukácsi nem hivatkozik. 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, például Martensnél (1979), ahol az állításokhoz kapcsolódó skálák nem kevesebb, mint 9 fokozatúak. A KER a vegyes megoldást nem tekinti külön kategóriának – nem is említi –, hanem besorolja a listás megoldások közé (PTMIK, 2002, p. 234), egyben gyengítve ezzel a skálás és listás megkülönböztetés erejét. A szakember lelkesedését lohasztja az is, hogy a listás pontok összesítésekor egy újabb, hosszú, sokfokozatú skála keletkezik éppúgy, mint a skálás megoldás alkalmazása során. Mivel az itt tárgyalt 7 szakdolgozati skála mindegyikére 0–4 pont adható, vagyis összesen 28 pont, az ELTE angoltanári összpontszám-skála is sokfokozatúnak tekinthető. Feltehetjük tehát a kérdést, hogy ha a szempontok és sávok száma pszichológiailag felülről korlátozott (Pollitt, 1991; Alderson et al., 1995; Luoma, 2004), mennyiben előnyösebb hosszú összpontskálát keletkeztető listákat alkalmazni, mint hasonlóan hosszú pontskálát eredményező analitikus skálákat. Mind a listás, mind a skálás megoldás, valamint a vegyes pontozási módszer használata után ugyanis olyan összpontszám-skálát kapunk, ahol a sikeresség minimumát zsűrizés (judgement) segítségével kell újra meghatározni, mivel a pontok az összpontszám-skálán már elválnak a külön skálákon még hozzájuk rendelt szintleírásoktól – így elveszhet a szintleírások nyújtotta értelmezhetőség, a kritériumfüggőség (criterion-referencing). Elmondható tehát, hogy a kétféle megközelítésmód inkább csak az értékelés eltérő technikai megoldásainak tekinthető.

*Az értékelők*

Az értékelő a mérés eredményére ható másik legfontosabb performanciahatás (tényező). A kutatások jellemző pedagógiai célja az értékelő eredményekre gyakorolt hatásának minimalizálása (pl. Kim, 2015, p. 239) – nem véletlenül, hiszen a mérési cél (kompetencia) mellett jelentkező performanciatényezők jellemző módon nemkívánatos módszertényezők: a mérés során az ér-

tékelő is nemkívánatos hatást fejt ki, mert szigorával, elnézőségével módosítja a ponteredményeket, pedig a mérésben nem is rá vagyunk kíváncsiak, hanem a szakdolgozatban megjelenő tudásra (kompetenciára).

A tanulmány szempontjából legfontosabb kutatási irány a hierarchikus értékelő modell (Hierarchical Rater Model, HRM) kifejlődéséhez kapcsolódó írások. Az értékelői hatás összetettségéről való gondolkodás három egymásból fejlődő elmélettel jellemezhető. Brown et al. (2005), Ducasse és Brown (2009), továbbá Eckes (2008) és Cumming et al. (2002) még az értékelői orientációt vizsgálták. Az orientációhoz kapcsolódó gondolatok precíz megfogalmazása DeCarlo (1998, 2005) szerint a szignáldetekt-elmélet (Signal Detection Theory, SDT) lett. E tanulmányokban – visszatekintve az SDT kezdeteire – DeCarlo megvilágítja, hogy az SDT viszont a hierarchikus értékelő modell (HRM) előképének tekinthető (Patz et al., 2002; DeCarlo, 2005).

A HRM jó magyarázat lehet eltérő értékelői tényezők sokrétűségére, beleértve az értékelői szerepeket is. DeCarlo et al. (2011) arra a következtetésre jutottak, hogy az íráskészség értékelésében eltérő szintek azonosíthatók. Az értékelők adta nyerspontértékek 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 hierarchikus értékelő modell tehát erősen emlékeztet a témavezető és bíráló eltérő értékelői szerepeire (helyzetére) a szakdolgozatok esetében. A hierarchikus modell hasznos, mert rávilágít arra, hogy az értékelők sem tekinthetők a performancia monolitikus tényezőjének. Összevetve az idegennyelv-tudás mérésével, megállapítható, hogy ha az első és második értékelő között nincs különbség a tekintetben, hogy mit tartanak feladatuknak, mire kell odafigyelniük stb. a nagymértékű értékelői egyetértés elvárható. Ha viszont a témavezető és bíráló szerepe eltér, értékelői konstruktumaik is eltérnek, és nem lehet arra számítani, hogy az értékelői egyetértés nagymértékű lesz. Hogy mire lehet számítani, mire nem, másodlagos – e helyett fontosabb lenne, hogy több kutatási beszámoló jelenjen meg skálák alkalmazása és értékelői gyakorlatok, eljárások vizsgálatáról (pl. Kiszely, 2012, 2019) a magyar nyelvű szakirodalomban.

## A KUTATÁS KÉRDÉSE

A jelen tanulmány arra a kutatási kérdésre keresi a választ, hogy mennyiben váltak be a szakdolgozati skálák rendszeres alkalmazásuk során a 2011 és 2019 közötti időszakban, a statisztika eszközrendszerének tükrében.

## A KUTATÁS MÓDSZERE

A módszerfejezet leírásába – az elemzési eljárások ismertetése mellett – beletartoznak az intézményi (tanárképzési program) keretek, ahol a kutatás végbement, a felhasznált adatok köre és maguknak a skáláknak a bemutatása, a vizsgálat szakaszainak bemutatása, valamint a mérés és érvényesítés elméleti háttere is.

### *A kutatás keretei*

A kutatás kereteit az ELTE Angol-Amerikai Intézete végzős, angoltanár-hallgatók (osztatlan tanárképzés, OTAK) szakdolgozatainak értékelése határozza meg. Ennek része volt 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 angoltanárszakos-hallgatók, akik szakpárosításuk másik szakja keretében készítették dolgozatukat.

### 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 értékelést tanúsítja, míg Dávid és Piniel (2018) elemzése csak a táblázatban szürkével jelölt, 2011 és 2013 közötti kezdeti időszakra terjedt ki. A skálák változtatás nélküli alkalmazása, továbbá a visszatérő értékelők pontozása mára lehetővé tette az adatcsomagok összevethetőségét (subset connection), amint azt a Facets szoftverbe épített ellenőrző algoritmus (Linacre, 2014, 2017) megnyugtató módon jelezte is. Hozzá kell tenni, hogy míg a tanulmányban feltett kutatási kérdések megválaszolásához megfelelők a 2011 és 2019 között felvett adatok, addig a 2020 és 2025 között gyűjtött adatok az értékelési rendszer változatlan alkalmazása mellett már új kérdéseket is felvetnek, mint például a COVID hatása, a dolgozatok elektronikus benyújtása és értékelése – azonban ez már egy másik kutatáshoz tartozik.

### A szakdolgozati skálák

A szakdolgozati skálák analitikusak, 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. Minden skálaponthoz deskriptor tartozik. A skálák az alárendelt szempontokkal együtt megtalálhatók a https://delp.elte.hu/otak_theses webhelyen, továbbá Dávid és Piniel (2018) B függelékeként is.

### Az adatelemzés módszerei

Az elemzések első szakasza a nyerspont-értékek konzisztenciájának (megbízhatóságának) vizsgálatát célozta a Chronbach Alpha kiszámításával (SPSS 27.0, IBM Corp., 2020). Ezt követte egy-

1 A mozaikszavak a szempontok angol nyelvű elnevezéseiből származnak és segítenek értelmezni a 3. táblázat sorait.

részt a szakdolgozati skálák megfelelése (beválása) mértékének vizsgálata, másrészt a kettős értékelési eljárás minőségének vizsgálata.

*A többdimenziós Rasch-elemzés (MFRM)*

Az elemzések további szakaszaiban 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, valamint azok szerepeit is. A dimenziókat a szociológia területéről származó facet theory alapján (Guttman & Greenbaum, 1998), Linacre (1989) faceteknek nevezte – innen a többdimenziós módszertan elnevezése is: Many-facet Rasch Measurement (MFRM). A probabilisztikus mérésre jellemző, logitban megadott mutatók elemzése a skálák megfelelésére vonatkozó kérdésre adott választ: a skálák mennyire illeszkednek a probabilisztikus mérési modellhez?

A skálák megfelelését az 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 ez utóbbi standardizált variánsa (outfit Z-érték) alapján. (Ezek a mutatók alább, a 3. táblázatban jelennek meg.) A point-biserial-típusú korrelációt a Linacre (2017) által javasolt point-measure (PtMeas) változat alapján vizsgáltam. A küszöbértékeket Linacre ajánlását módosítva alkalmaztam. A küszöbértékek azért fontosak, mert az ezek feltetti megmagyarázatlan szórás már nem elfogadható. Feltételeztem, hogy az illeszkedési értékek normál eloszlásúak, ezért a skála, ill. skálapont (pontérték és deskriptor) illeszkedési mutatója az átlag felett két szórásegységen (standard deviation, SD) túl már extrém, elfogadhatatlan értéknek tekinthető.

## A vizsgálat szakaszai

Az elemzések korrelációk sorával, pontátlagok összevetésével, az egyetértés vizsgálatával stb. folytatódtak az alábbi szakaszok szerint:

1. szakasz: A vizsgált pontértékek megbízhatósága (konzisztencája).
2. szakasz: A vizsgálat a témavezető és a bíráló összesített pontszámainak korreláltatása és a Fleiss-féle Kappa egyetértési mutató kiszámítása.
3. szakasz: A korrelációk következő csoportja, mind a témavezető, mind pedig a bíráló esetében az összesített pontszám és a javasolt érdemjegy korrelációja, egyben a pontszám-érdemjegy konverzió mérlegre tétele.
4. szakasz: A kutatás azt is vizsgálta – mint a korrelációk harmadik csoportját – hogy a témavezető, illetve a bíráló mennyiben hatott sajátos módon a szakdolgozat végső eredményére, továbbá a témavezető és a bíráló egyeztetésének hatása is itt volt vizsgálható. Hozzáteszem, hogy a korrelációs érték nem értelmezhető ok-okozati viszonyként, hanem csak adott hasonlóságként vagy együttállásként.

A korrelációkat kiegészítették párosított t-próbák az átlageltérések vizsgálatára. Az összes elemzési szakasz (1–4) áttekinthetősége szempontjából hasznos azok táblázatos ábrázolása (2. táblázat).

**2. táblázat**

*Az elvégzett konzisztencia- és korrelációs vizsgálatok táblázatos áttekintése*

| Elemzési szakasz | Az összevetés alapja | |
|---|---|---|
| 1 | Az összes (310 ×2 ×7) pontérték konzisztenciája a mérésben | |
| 2 | Témavezető és a bíráló összesített pontjainak korrelációja | |
| 3 | Témavezető összesített pontjai és javasolt korrelációja | Bíráló összesített pontjai és javasolt jegyeinek korrelációja |
| 4 | Témavezető összesített pontjai és a végső jegyeinek korrelációja | Bíráló összesített pontjai és a végső jegyeinek korrelációja |

### Az érvényesítés elméleti háttere

A kutatás bevett érvényesítési eljárást követ, amelyben először az összes válaszadat konzisztenciáját (megbízhatóságát) és az értékelői megbízhatóságot szokás vizsgálni (Brennan, 2006, p. 5), amit később követhet az érvényesség megállapítását célzó további munka, ha a mérési eszközrendszer (skálák, értékelők) megbízhatósága kielégítő. E klasszikus eljárás szerint amennyiben a megbízhatóság nem kielégítő, az eredmények érvényességét sem lenne érdemes tovább vizsgálni, ugyanis megbízhatósága a válaszadatoknak van, érvényessége az eredményeknek és azok következményeinek kell legyen (Messick, 1995). Az itt közölt mérések csak az érvényesítést megalapozó bizonyítékoknak tekinthetők.

## A KUTATÁS EREDMÉNYEI

### A pontozás konzisztenciája, megbízhatósága

A Chronbach Alpha, összesen 310 dolgozat adatainak betáplálásával elérte a 0,914 értéket, dolgozatonként 14 pontérték alapján, mivel az elemzés a két értékelő 7-7 pontjára kiterjedt. A konzisztenciát a pontok érdemjegyekre váltása fenntartja és tovább vetíti. A bíráló pontjainak és a konverziós táblázat szerinti érdemjegyek korrelációja r=0,901 (N = 309), amelytől kicsit elmarad a témavezető pontjainak és a konverziós tábla szerinti érdemjegyeinek korrelációja: r = 0,883 (N = 304).

### A skálák megfelelése, beválása

Általánosságban megállapítható, hogy a skálák beváltották a hozzájuk fűzött elvárásokat (3. táblázat) – 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. Kivételt képez a *Formai követelmények* (FoR)-skála, amely legfeljebb marginálisan fogadható el. A skála infit meansquare értéke magas ugyan, de még épp megfelelő (1,29 a táblázatban vastagon szedve), viszont az outfit-érték már túlmegy a 2

szórásegységen (SD) és extrémnek tekinthető. Mivel az outfit súlyozatlan érték, érzékeny a nem várt, kiugró (outlier) válaszokra, esetleg valamely más skála miatt. Ha az infit és outfit standardizált (Z) értékekeit vizsgáljuk, azok már nem jeleznek túl sok megmagyarázatlan szórást, már nem lépik túl a határértékeket. A 3. táblázatban további tájékoztatás érhető el a skálák megfeleléséről: a nyerspontok átlaga (Nyers átl., 0 és 4 pont között), a logit nehézségi érték (Nehézség), valamint a point-measure-korreláció megfigyelt (PtMeas) és elvárt (Ptelvárt) értékéről is.

**3. táblázat**

*Az ELTE Angol-Amerikai Intézet angoltanári skálák megfelelésének összefoglaló táblázata*

| Skálák | Nyers átl. | Nehézség | Hibaérték | InfitMS | InfitZ | OutfitMS | OutfitZ | Diszkrim. | PtMeas. | PtElvárt |
|---|---|---|---|---|---|---|---|---|---|---|
| QuWr | 3,32 | 0,79 | 0,07 | 0,86 | -2,54 | 0,83 | -2,81 | 1,17 | 0,72 | 0,67 |
| IntF | 3,33 | 0,61 | 0,07 | 0,96 | -0,76 | 0,93 | -1,07 | 1,06 | 0,68 | 0,67 |
| ReMePr | 3,18 | 0,32 | 0,07 | 0,95 | -0,78 | 0,92 | -1.2 | 1,05 | 0,72 | 0,71 |
| QuEL | 3,1 | -0,17 | 0,07 | 1,12 | 2,02 | 1,12 | 2.07 | 0,86 | 0,62 | 0,66 |
| ThEB | 3,35 | -0,42 | 0,08 | 0,91 | -1,58 | 0,93 | -1.04 | 1,09 | 0,69 | 0,66 |
| FoR | **3,32** | **-0,48** | 0,07 | 1,23 | 3,61 | **1,29** | 3.92 | **0,72** | **0,60** | **0,68** |
| Ind | 3,55 | -0,66 | 0,08 | 0,9 | -1,36 | 0,87 | -1.25 | 1,06 | 0,67 | 0,65 |
| Átlagok | 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 *Formai követelmények*-skála egyéb tekintetben, a klasszikus tesztelmélettől vett mutatók szerint sem jeleskedik. Ez a skála egyben a leggyengébben diszkrimináló (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 a többi skála között (0,60), továbbá elvárt értéke magasabb (0,68), mint a mért érték. A mért és elvárt értékek közötti eltérés is ennél a skálánál a legnagyobb. Van további fontos és ellentmondásos mutató is: ez a skála a második „legkönnyebb" skála, átlagos logit nehézségi értéke (-0,48) alapján, azonban a nyerspont-átlagok 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 a 4. 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) által javasolt értékhez (1,4) képest, de a Dávid és Piniel (2018) által közölt korábbi értékeknél mindenképpen jobb. A skálapontok outfit-értékei is magasak; az 1. és 3. skálapontok tekintetében az elfogadhatóság határán mozognak.

**4. 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.

## További korrelációs összefüggések és az átlagok vizsgálata

A témavezetői és bírálói összpontszámok Pearson korrelációja közepesnek minősíthető és szignifikáns (r = 0,591, df = 303, p. < 0,01). Az általános egyetértés mértéke a Fleiss-féle multirater Kappával jellemezhető, melynek értéke 0,181 volt – meglehetősen alacsony –, de a konzisztenciához hasonlóan szignifikáns (p.< 0,01). Mindez arra utal, hogy van különbség a témavezetők és bírálók pontozása között.

További érdekes összefüggés még a témavezető és a bíráló pontjai összegének és a végső érdemjegynek korrelációs kapcsolata, mert – mint ahogy a bevezetésben is szerepel –, a témavezető és bíráló által javasolt jegyek egy érdemjegynyi eltérése esetén felkérést kapnak a koordinátortól, hogy döntsék el a végső jegyet. Ugyanazzal a végső érdemjeggyel korrelál tehát a témavezető pontszáma (r = 0,681) és a bírálóé (r = 0,758). Mint látható, a bírálóhoz kötött korrelációs érték a magasabb ugyan, de a különbség nem jelentős.

A magas belső konzisztencia mellett jelentkező különbségekre a témavezetői és bírálói összpontszámátlagok és javasolt jegyátlagok is rámutatnak. Az átlagok azt mutatják, hogy a bírálók szigorúbbak, a témavezetők pedig elnézőbbek (5. táblázat). Ez az eltérés önmagában nem is lenne meglepő, ha a bírálók és témavezetők elkülönült csoportot alkotnának.

**5. 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 |

### Fontos következtetés

Az eltérések azonban mégis azért különösen érdekesek, mert túlnyomó részt ugyanazok a kollégák pontoznak – vagy témavezetőként, vagy bírálóként –, jellemzően ugyanabban a félévben. Az adattábla szerint a témavezetők 82%-a (31/38 fő) egyben bíráló is volt, míg a bírálók 87% -a (27/31) egyúttal témavezető is, vagy – megfordítva a fenti megállapításokat – az értékelők egy kisebb része volt csak témavezető (18%), vagy csak bíráló (13%). A pontátlagok különbségei, azok egyszerű összegeit tekintve ($t = 6{,}368$, $df = 302$, $p < 0{,}01$), mind pedig a javasolt jegyek tekintetében ($t = 4{,}822$, $df = 302$, $p < 0{,}01$) szignifikánsak. A megfigyelt különbségek érvényességét megerősíti tehát a kutatás közel teljes mértékben keresztelt szerkezete. Valószínűtlen, hogy például a témavezetők magasabb pontátlagát csak eme csoport elnézősége magyarázná.

## Összegzés

A skálák megbízható alkalmazása azt jelentette, hogy megállták a helyüket a szakdolgozatok értékelésében. Az értékelés dolgozatonkénti 2×7 (14) pontértékére szükség van, mert csupán egy értékelő alkalmazása esetén a Cronbach Alpha 0,777 lenne, ha csak a bíráló értékelne a 7 szempont alapján, míg 0,775 ha csak a témavezető tenné ugyanazt. A skálák megbízható alkalmazása azt is jelentette, hogy a ponteredmények alapján meghozott döntések érvényességének további vizsgálata lehet a jövőbeli kutatás egy iránya.

Megállapítható az is, hogy a skálák többségének megfelelősége jó. A *Formai követelmények*-skála pontozása mögött álló gondolkodás elsősorban figyelmet, további tréninget és kutatást igényel, melyben kvalitatív módszerekkel választ kaphatunk arra a kérdésre, mi okozza a *Formai követelmények*-skála itt megfigyelt illeszkedési problémáját. Ha a tréning nem vezetne megfelelő eredményre, a skála elemzésekre épülő módosítása is a lehetséges forgatókönyvek egyike. Ebben az esetben az új skálát moderálni és pretesztelni kell. Bár Dávid és Piniel (2018) korábbi, 78 dolgozatra épülő elemzésükben az illeszkedési problémát már azonosították, az javuló tendenciát mutat az újabb adatok és a 310 dolgozatra épülő 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 nem abszolút prioritás.

További összegzésként, és egyben kitekintésként elmondható, hogy a bíráló és témavezető közötti egyetértés (vagy egyet nem értés) Fleiss-féle Kappával mért alacsony értéke és a bírálói és témavezetői kör igen jelentős átfedése között ellentmondás feszül – mintha az értékelők önmagukkal nem értenének egyet. Az értékelők egyetértési mutatójának alacsony volta arra enged következtetni, hogy a témavezető és bíráló nem azonos szemlélettel értékeli a dolgozatokat. Az ellentmondásból adódó kérdést terjedelmi okokból egy másik tanulmányban válaszolom meg.

### Köszönetnyilvánítás

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

## IRODALOM


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

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

Bramley, T., & Dhawan, V. (2011). Investigating and reporting information about marker reliability in high-stakes external school examinations [Conference presentation]. *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 R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 1–16). American Council on Education/Praeger.

Brown, A., Iwashita, N., & McNamara, T. (2005). An examination of rater orientations and test-taker performance on English-for-academic-purposes speaking tasks (TOEFL Monograph No. 29). *ETS Research Report Series*, 2005(1), i–157. https://doi.org/10.1002/j.2333-8504.2005.tb01982.x

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. https://doi.org/10.1111/1540-4781.00137

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. https://www.columbia.edu/~ld208/psymeth98.pdf

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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1177/0265532209104669

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

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

Hallinger, P. (2010). A review of three decades of doctoral studies using the principal instructional management rating scale. *Educational Administration Quarterly, 47*(2), 271–306. https://doi.org/10.1177/0013161X10383412

IBM Corp. (2020). *IBM SPSS Statistics for Windows (Version 27.0)* [Computer software].

Kim, H. J. (2015). A qualitative analysis of rater behavior on an L2 speaking assessment. *Language Assessment Quarterly, 12*(3), 239–261. https://doi.org/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.* [Doctoral dissertation, University of Toronto]. https://hdl.handle.net/1807/24786

Kim, Y.-H. (2011). Diagnosing EAP writing ability using the reduced reparameterized unified model. *Language Testing, 28*(4), 509–541. https://doi.org/10.1177/0265532211400860

Kiszely, Z. (2012). Egy anglisztika szakon használt szakdolgozati értékelési skála megújításának alapelvei és használatának első eredményei. In C. Sárdi (Ed.), *A felsőoktatás-pedagógia kihívásai a 21. században* (pp. 221-234). 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/modernnyelvok/issue/view/136/60

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

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

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

Lukácsi, Z. (2020). Developing a level-specific checklist for assessing EFL writing. *Language Testing, 38*(1), 86–105. https://doi.org/10.1177/0265532220916703

Luoma, S. (2004). *Assessing speaking*. Cambridge University Press.

Martens, F. L. (1979). A scale for measuring attitude toward physical education in the elementary school. *The Journal of Experim*ental Education, 47(3), 239–247. http://www.jstor.org/stable/20151282

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

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

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. https://doi.org/10.3102/10769986027004341

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

PTMIK. (2002). *Közös Európai Referenciakeret: Nyelvtanulás, nyelvtanítás, értékelés*. Pedagógus-továbbképzési Módszertani és Információs Központ. [fordítás]

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. https://doi.org/10.1016/j.jslw.2023.101007

Sideridis, G., & Padeliadu, S. (2011). Creating a brief rating scale for the assessment of learning disabilities. *Journal of Learning Disabilities, 46*(2), 115–132. https://doi.org/10.1177/0022219411407924

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

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press.

Williams, L., & Kemp, S. (2018). Independent markers of master's theses show low levels of agreement. *Assessment & Evaluation in Higher Education, 44*(5), 764–771. https://doi.org/10.1080/02602938.2018.1535052

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. https://doi.org/10.1177/0734282912456969