Beérkezett: 2025. 11. 02. Elfogadva: 2025. 12. 29. Megjelent: 2026. 02. 28.

DOI: 10.21030/anyp.2026.1.2

Juhász Milán

A fogalmazásértékelés megbízhatóságának empirikus vizsgálata két értékelési modell összehasonlításával

A tanulmány a szövegalkotás értékelésének megbízhatóságát vizsgálja a magyar nyelv és irodalom középszintű érettségi vizsga kontextusában, empirikus adatokra támaszkodva. A kutatás célja annak feltárása volt, hogy a magyar nyelv és irodalom középszintű érettségi vizsgán 2024-től alkalmazott javítási-értékelési útmutató milyen mértékben biztosít objektív, következetes és az értékelők között egységesen értelmezett pontozási keretet a tanulói szövegek értékeléséhez, valamint egy alternatív, analitikus értékelési szempontrendszer alkalmazása képes-e növelni az értékelők közötti egyetértést. A vizsgálatban negyven, érettségiztetési gyakorlattal rendelkező magyar nyelv és irodalom szakos pedagógus vett részt, akik egy tanulói irodalmi elemző esszét pontoztak két eltérő értékelési rendszer alapján. Az elemzés az értékelők közötti egyetértést és a mérőeszközök belső konzisztenciáját több megbízhatósági mutató segítségével vizsgálja, emellett kitér az egyes értékelési szempontok differenciáló erejére és az értékelői mintázatok eltéréseire is. Az eredmények alapján az érettségi vizsgán alkalmazott szempontrendszerben alacsonyabb az értékelők közötti egyetértés, miközben a szempontok értelmezése számottevő variabilitást mutat. Ezzel szemben az alternatív, analitikus értékelési modell alkalmazásával magasabb egyetértési szinteket és stabilabb értékelési struktúrát lehet elérni. Az analitikus szempontsor részletesebb teljesítményleírásai hozzájárulnak az értékelői döntések következetesebb alkalmazásához, ugyanakkor az eredmények arra is rámutatnak, hogy az értékelési dimenziók további pontosítása indokolt lenne. A tanulmány következtetései alátámasztják az értékelési kritériumok világosabb meghatározásának és a standardizáltabb fogalmazásértékelési gyakorlat kialakításának a szükségességét.

Bevezetés

A szövegalkotás értékelése a pedagógiai gyakorlat egyik legösszetettebb és legérzékenyebb területe. Bár a tanulói íráskészség fejlesztése a magyar nyelv és irodalom oktatásának egyik alapvető célja, a fogalmazások értékelésével kapcsolatos kutatások száma a hazai szakirodalomban viszonylag alacsony. Ennek oka elsősorban a jelenség komplexitásában rejlik (P. Tóth 2025). A szövegalkotás minősége több, egymással szorosan összefüggő dimenzió mentén értékelhető, amelyek együttes vizsgálata komoly módszertani kihívást jelent. Az értékelési folyamatot tovább nehezíti, hogy a bírálói ítéletek gyakran szubjektív elemeket is tartalmaznak, az értékelés megbízhatósága pedig több egyéb tényezőtől is függhet.

A jelen tanulmány célja, hogy empirikus adatokon keresztül vizsgálja a szövegalkotás értékelésének megbízhatóságát és az értékelői ítéletek konzisztenciáját a középszintű érettségi vizsga kontextusában. A fogalmazásértékelés komplex folyamatát számos tényező befolyásolja, például az értékelési skálapontok definiálása (Nagy 2013), a pedagógus szigorúsága (Eckes 2008), a bíráló gondolkodásának kulturális meghatározottsága, személyiségjegyei, tulajdonságai (Szilassy 2012; Fülöp 2017), a pedagógus szövegalkotásra vonatkozó tudása, elvárásai (Horváth 1998; Tóth 2008; Kerner 2009), de a tanítási tapasztalat hossza is meghatározó lehet (Juhász 2021). Mindezen tényezők közül a skálapontok és az alkalmazott értékelési szempontsorok azok, amelyek a legközvetlenebbül megragadhatók és vizsgálhatók. A tanulmány

ezért elsősorban az értékelési szempontrendszerekre koncentrál, és azt elemzi, hogy a 2024-től alkalmazott érettségi szempontsor milyen mértékben képes biztosítani az objektivitást és a bírálók közötti egységes értelmezést. Emellett a vizsgálat arra is választ keres, hogy egy alternatív, analitikus mérőeszköz alkalmazása növelheti-e az értékelések megbízhatóságát, csökkentheti-e az egyéni értelmezésből fakadó eltéréseket, és hozzájárulhat-e a szövegalkotási kompetencia árnyaltabb, kiegyensúlyozottabb méréséhez. A tanulmány egy korábbi, a fogalmazásértékelés megbízhatóságát feltáró kutatás empirikus eredményeit mutatja be (Juhász 2025).

A középszintű érettségi értékelési szempontsora

A 2020-ban hatályba lépett legújabb Nemzeti alaptanterv és az ahhoz illeszkedő kerettantervek bevezetése jelentős változásokat eredményezett a magyar nyelv és irodalom tantárgy érettségi vizsgarendszerében. Ezen szabályozók mentén dolgozták ki az új, részletes érettségi vizsgakövetelményeket, amelyek az oktatási tartalmak, a kompetenciafejlesztési célok és az értékelési kritériumok szempontjából is igazodnak az új tantervi irányelvekhez (1).

A magyar nyelv és irodalom érettségi vizsgakövetelményei átfogó módosításokon mentek keresztül mind a vizsgafeladatok, mind az értékelési rendszer vonatkozásában (a részletes változásokról bővebben Karkó 2022). A középszintű írásbeli vizsga első vizsgarésze továbbra is kilencven percet biztosít a vizsgázók számára, a korábbi években alkalmazott rövid szövegalkotási feladat azonban kikerült ebből a vizsgarészből. Az új vizsgaszerkezet első része egy szövegértési-nyelvi feladatsort (maximum 40 pont) és egy irodalmi feladatlapot (maximum 20 pont) tartalmaz, amelyek egymástól elkülönülten mérik a tanulók tudását és képességeit.

Az írásbeli vizsga második részében a vizsgázóknak továbbra is egy szövegalkotási feladatot kell megoldaniuk, 500–800 szó terjedelemben. A tanulók két téma közül választhatnak: egy irodalmi mű értelmezését végzik el, vagy egy témakifejtő dolgozatot/esszét alkotnak meg az adott témáról. A feladathoz kapcsolódó javítási-értékelési útmutató struktúrája változatlanul két fő részből áll: az általános értékelési szempontokból, amelyek mindkét feladattípusra érvényesek, valamint az egyes feladatokhoz specifikusan rendelt értékelési szempontokból. Az értékelési rendszer – főképp a pontszámok tekintetében – azonban jelentős módosításokon ment keresztül: a pontozás rendszere átalakult, miközben a teljesítménytartományokhoz kapcsolódó leírásokat csupán minimálisan módosították. Az új értékelési rendszerben az írásmű tartalma vált dominánssá, amelyre a javító tanár legfeljebb harminc pontot adhat, a szemponton belül a teljesítményt öt sávba sorolhatja. A nyelvi megformálás és stílus az új rendszerben veszít a súlyából: a nyelvhasználat, stílus szempontjára az előző értékelési modellben maximum 10 pontot szerezhettek a vizsgázók, míg az új modellben a nyelvi igényességért maximum 5 pont jár. A szövegszerkezet szempontja ugyanakkor megmaradt 5 pontos kategóriaként. Az íráskép és a helyesírás értékelésén belül jelentős szemléletváltás következett be: míg eddig a jó helyesírást jutalmazták, azaz pontot lehetett szerezni vele, a 2024-től alkalmazott értékelési modellben a hibákat büntetik, ezekért pontlevonás jár. A helyesírásért maximum nyolc pont, az írásképért pedig két pont vonható le.

A tartalomra adható pontszám a szövegalkotási feladat maximum pontszámának jelentős részét teszi ki: az elérhető negyven pontból harminc pontot ér. Ehhez viszonyítva a szöveg szerkezetének és nyelvi igényességének a minősítése jóval alacsonyabb pontszámokkal valósul meg a rendszerben, ami felveti a kérdést, hogy ezen szempontok megfelelő súllyal érvényesülnek-e az értékelés során, és hogy mennyire tudja kiegyensúlyozottan mérni a szövegalkotási kompetencia fejlettségét az új rendszer. A szempontrendszer egyik legkritikusabb pontja az, hogy a tartalmi teljesítmény ugyan differenciált, de továbbra is a szubjektivitást sem kizáró sávos besorolással értékel, sávonként 5 ponttal. Az értékelési sávokban szereplő kritériumok,

mint a „kifejtett, indokolt, tárgyszerű állítások” vagy a „megfelelő tájékozottság” túlságosan tág fogalmi keretet alkotnak, amelyeket a javító tanárok eltérően értelmezhetnek (Szentgyörgyi 2017). A szubjektivitás kockázatát tovább növelheti, hogy az útmutató csupán szöveges leírást tartalmaz, kvantitatív indikátorok vagy normatív mintaszövegek nélkül. A korábbi, 2024-ig használt értékelési szempontsor megbízhatósági elemzésének eredményei is azt mutatták, hogy mindezek miatt ugyanarra a szövegre a független bírálók eltérő pontszámokat adhatnak, és ez az érettségi vizsga mérési objektivitását csökkentheti (Juhász 2022).

Az analitikusabb értékelési mérőeszközök, például a különböző tételek mentén meghatározott részletesebb rubrikarendszer, csökkenthetik az eltéréseket az értékelésben. Ha megfigyeljük a nemzetközi példákat, látható, hogy a részletesebben kidolgozott értékelési skálák alkalmazása jobban hozzájárul a konzisztensebb pontozáshoz, így ezek képesek növelni az értékelés objektivitását (Yaqub et al. 2016; Aslim Yetis 2019), emellett a tanulók szövegalkotási képességének a fejlesztésében is hatékonyabbak, hiszen képesek strukturált visszacsatolást és világos teljesítményelvárásokat adni a tanulóknak az önszabályozott tanuláshoz (Li 2022). A nemzetközi szakirodalom, valamint a hazai fogalmazáskutatások eredményei szolgáltak kiindulópontként a következő részben ismertetett alternatív értékelési szempontsor kidolgozásához.

Egy alternatív mérőeszköz bemutatása

A következőkben bemutatott alternatív mérőeszköz tíz különálló értékelési kategóriát tartalmaz, és minden egyes kritériumot egy ötfokú skálán osztályoz. Az ilyen típusú analitikus értékelési keretrendszer legfontosabb előnye, hogy redukálja az értékelők közötti különbséget azáltal, hogy pontosan operacionalizált teljesítményszinteket határoz meg, ezzel minimalizálva az értelmezési eltéréseket (Aslim Yetis 2019). A szempontsor minden egyes értékelési kategóriára ötpontos skálát alkalmaz, amely jelentősen szűkíti a pontozási varianciát, így csökkenti az értékelési inkonzisztenciák lehetőségét. Minden szempont egyenlő súllyal szerepel a rendszerben, azaz minden kritérium maximálisan 5 pontot érhet. Ez biztosítja, hogy a különböző szempontok azonos mértékben járuljanak hozzá a végső értékeléshez, elkerülve a jelenlegi érettségi rendszerben tapasztalható kiegyensúlyozatlanságot, amelyben a tartalmi kritériumok dominanciája miatt más szempontok háttérbe szorulnak.

Az alternatív mérőeszköz további lényegi előnye az, hogy minden teljesítményszinthez igyekszik pontosan meghatározott kritériumokat rendelni, így az értékelők számára egyértelműbbé válik, hogy egy adott pontszám milyen konkrét szövegjellemzőket feltételez. Az alternatív szempontsor felépítését az 1. táblázat szemlélteti. A szempontok meghatározásánál és a kritériumrendszer kidolgozásánál megjelennek a hazai fogalmazáskutatásokban használt értékelési rendszerek (például Orosz 1972; Kádárné Fülöp 1990; Horváth 1998; Molnár 2000; Nagy 2009; Szilassy 2012; Fülöp 2017), valamint a nemzetközi értékelési modellek struktúrái. Emellett referenciapontként szolgált Nagy Zsuzsanna kilenc analitikus és egy holisztikus szempontot integráló saját fejlesztésű értékelési eszköze is (Nagy 2013: 160), amely komplex módon közelíti meg a szövegalkotási teljesítmény mérését. Ezek együttes figyelembevételével olyan értékelési szempontsor kialakítása volt a cél, amely empirikusan megalapozott, és illeszkedik a hazai és a nemzetközi értékelési gyakorlatokhoz.

1. táblázat
Az alternatív analitikus értékelési szempontsor felépítése

Szempont	5 pont	4 pont	3 pont	2 pont	1 pont
Tartalom	Teljes, releváns érvek, következetes és mélyreható kifejtés. Minden állítás alá van támasztva példával.	Többnyire releváns érvek, kisebb hiányosságokkal. Egy-két állítás nincs teljesen alátámasztva.	Közepesen kidolgozott, néhány kulcselem hiányzik. Például egy fontos érv kifejtetlen marad.	Felületes, lényegi érvek hiányoznak. Csak néhány gondolat van kifejtve, nem megfelelő mélységben.	Nagyon hiányos, alig van tartalom. Az érvek nincsenek kifejtve.
Feladattartás – szövegtípus	Teljes mértékben megfelel a műfaji követelményeknek.	Kisebb eltérések vannak, de alapvetően a szövegtípusnak megfelelő a megoldás.	A megoldás csak részben felel meg a szövegtípusnak, néhol eltérések vannak.	A megoldás nem illeszkedik teljesen a szövegtípushoz.	Nem az elvárt műfajban íródott. Például esszé helyett csak jegyzetek, vázlatpontok szerepelnek.
Feladattartás – hangnem	A stílus és a hangnem következetesen adekvát, igazodik a szövegtípushoz (például hivatalos vagy személyes hangnem).	Apróbb stilisztikai eltérések vannak, de a szöveg összességében illeszkedik a műfajhoz.	Részben megfelelő hangnem, néhol nem illeszkedik a szöveg típusához. Például túl közvetlen kifejezések.	Gyakran eltér a megfelelő hangnemtől. Például irodalmi esszében szleng használata.	Teljesen inadekvát hangnem. Például irodalmi esszében trágárság vagy túlzott lazaság fordul elő.
Szerkezet és kidolgozás	Logikus, jól strukturált, összefüggő szöveg. Minden bekezdés világosan kapcsolódik az előzőhöz. Megvannak a főbb szerkezeti egységek. Megfelelően tagolt.	Többnyire koherens, kisebb logikai hiányosságokkal. Egyes bekezdések kevésbé kapcsolódnak egymáshoz.	Részben koherens, néhol szervezetlen. Egyes gondolatok nem állnak logikus sorrendben. Nem tagolt.	Nehezen követhető, széttöredezett gondolatmenet. A bekezdések között nincs összefüggés. Hiányzó szerkezeti egységek (például bevezetés).	Kaotikus, következetlen gondolatok. Nincs világos szerkezet.
Stílus	Gazdag, változatos szókincs, világos kifejezésmód, szakkifejezések alkalmazása. Szinonimák és retorikai eszközök tudatos használata.	Többnyire választékos, kisebb stílushibákkal. Előfordul néhány ismétlés vagy pontatlan kifejezés.	Egyszerű, de érthető nyelvezet. A szókincs nem túl gazdag, de nem is túl primitív.	Pontatlan vagy nem megfelelő stílus. Például gyakori szóismétlések és közhelyek használata.	Gyenge, igénytelen nyelvhasználat. Nagyon szegényes szóhasználat.
Érthetőség	Teljes mértékben érthető, világos gondolatmenet. Az olvasó számára egyértelmű minden következtetés.	Néhol nehezen követhető, de az összkép érthető. Egyes mondatok túlbonyolítottak lehetnek.	Részben érthető, kisebb zavarokkal. Előfordulhatnak félreérthető mondatok.	Több helyen érthetetlen, nehezen követhető gondolatmenet.	Érthetetlen, zavaros. Az olvasó nem tudja követni a gondolatmenetet.
Nyelvhelyesség	Hibátlan mondatszerkesztés és nyelvhasználat. Például alárendelt és mellérendelt mondatok megfelelő használata.	Kevés, nem súlyos nyelvtani hiba. Például néhány helytelen egyeztetés, ami az érthetőséget nem zavarja.	Több kisebb nyelvtani hiba, de az érthetőséget nem veszélyezteti. Például néhány elhibázott igeidő.	Gyakori nyelvtani hibák, amelyek rontják az érthetőséget. Például alany-állítmány egyeztetési hibák.	Súlyos, érthetetlenné tevő hibák. Például folyamatos alaktani és mondatszerkesztési hibák.

Szempont	5 pont	4 pont	3 pont	2 pont	1 pont
Helyesírás	0–5 helyesírási hiba.	5,5–8,5 helyesírási hiba.	9–12 helyesírási hiba.	12,5–15,5 helyesírási hiba.	16 vagy több helyesírási hiba.
Külső megjelenés	Olvasható, esztétikus szöveg. Rendezett margók, egységes betűméret. Esztétikus javítás.	Többnyire rendezett, kisebb formai hiányosságokkal. Például egyenetlen bekezdéshossz.	Helyenként rendezetlen. Egyes sorok nehezen olvashatók.	Nehezen olvasható, rendezetlen. Rossz sortörések, rendezetlen bekezdések. Az ékezetek nem felismerhetők.	Olvashatatlan, nagyon rendezetlen. Például túlzsúfolt, nem esztétikus javítás.
Összbenyomás	Nagyon jól kidolgozott, kiforrott szöveg. A tartalom, a szerkezet és a nyelvi minőség egyaránt magas színvonalú.	Jó minőségű, de kisebb hiányosságokkal. Fejleszthető egyes részekben.	Elfogadható, néhány fejlesztendő területtel. Általánosan jó, de hiányosságai vannak.	Gyenge minőségű, sok hiányossággal. Átfogó fejlesztés szükséges.	Nem megfelelő, alig kidolgozott. Nem felel meg az elvárásoknak.

Az analitikus rendszer lehetőséget biztosít arra, hogy az értékelés részletesebb és következetesebb legyen, ezáltal minimalizálja az értékelők közötti szubjektív eltéréseket is. Az érettségi vizsgán alkalmazott rendszerrel szemben minden szempontot azonos súllyal kezel, így kiegyensúlyozottabb értékelési struktúrát biztosít. A helyesírás szintén önálló, ötfokú skálán értékelt dimenzióként szerepel, az értékelés azonban nem a középszintű érettségi vizsga hivatalos, levonásos pontozási rendszerének a reprodukcióját célozza, hanem kutatási célú, standardizált mérési megoldásként funkcionál. A hibák azonosítása a magyar helyesírás szabályai alapján történik, a hibák típus szerinti súlyozása azonban nem valósul meg: minden hiba azonos súllyal jelenik meg a módszertani egységesség érdekében, leszámítva a központozási hibákat, amelyek fél pont levonással járnak. A helyesírás ilyen módon történő operacionalizálása nem normatív javaslatként, hanem mérési konstrukcióként értelmezendő. Ahogy az eredményeknél majd látható, a kutatás egyik célja annak vizsgálata volt, hogy egy transzparens, analitikusan tagolt skála miként hat az értékelők közötti egyetértésre. Az eredmények alapján a helyesírás dimenziójában az alternatív szempontrendszer alkalmazása mellett kiemelkedően magas fokú egyezés volt tapasztalható az értékelők között.

Kutatási kérdések, hipotézisek, minta és módszertan

A kutatás célja az volt, hogy empirikus vizsgálat keretében elemezze a középszintű érettségi vizsgán a 2024-es évtől alkalmazott értékelési szempontrendszer megbízhatóságát, valamint a fentiekben bemutatott alternatív, analitikus értékelési keretrendszer hatékonyságát az értékelők közötti eltérésekre és az értékelési szempontok differenciáló erejére fókuszálva. A vizsgálatban egy középiskolai tanuló irodalmi elemző esszéjének két eltérő értékelési keretrendszer szerinti bírálatát végezte el összesen negyven középiskolai magyar nyelv és irodalom szakos pedagógus, akik már rendelkeznek érettségiztetői gyakorlattal. A minta kialakításánál cél volt az intézménytípus és a szakmai tapasztalat szerinti egyensúly fenntartása. A pedagógusok fele (20 fő) gimnáziumban, míg a másik fele (20 fő) egyéb intézménytípusban (például szakiskolában) tanít. A gyakorlati idő szerinti megoszlás szintén kiegyensúlyozott: a minta fele pályakezdő pedagógusnak tekinthető (1–5 év tanítási tapasztalattal), míg a másik fele legalább 5 év szakmai gyakorlattal rendelkező, tapasztaltabb tanár. A minősítésre kiválasztott szöveg egy 11. évfolyamos tanuló irodalmi esszéje, amely Karinthy Frigyes *Az olasz fagylaltos* című novellájának értelmezésére készült. Az írásbeli feladat a 2024-es középszintű magyar nyelv és irodalom érettségi vizsga feladatlapjából származik (2: 19).

A bíráló pedagógusok először a középszintű érettségi vizsgán alkalmazott központi javítási-értékelési útmutató alapján javították és pontozták a szöveget (3), majd négy hónap múlva ugyanazt a tanulói fogalmazást az

alternatív, analitikus értékelési szempontsor alkalmazásával minősítették újra. A vizsgálat összeveti a két értékelési rendszer alkalmazásával elért eredményeket, kitérve az egyes szempontokra kapott pontszámok közötti variancia elemzésére is. A vizsgálat elsődleges célja az volt, hogy választ adjon arra az alapvető kérdésre, hogy mennyiben tekinthető megbízhatónak a jelenlegi érettségi vizsgán alkalmazott értékelési szempontrendszer, illetve hogy az alternatív mérőeszköz képes-e csökkenteni az értékelők közötti eltéréseket. A kutatási kérdések, amelyek szervezték a vizsgálatot, a következők:

– Milyen mértékű az értékelők közötti egyetértés az érettségi vizsgán alkalmazott szempontrendszer szerinti javítás során?
– Hogyan változik az értékelési konzisztencia, ha az alternatív, analitikus értékelési szempontrendszert alkalmazzák a bírálók?
– Milyen különbségek figyelhetők meg az egyes pedagógusok értékelési mintázataiban, és ezek milyen mértékben befolyásolják az értékelési folyamat megbízhatóságát?
– Milyen kapcsolat mutatható ki az analitikus szempontok és az összbenyomás között az alternatív szempontsor esetén?

A kapcsolódó hipotézisek a következők:

(1) Az érettségi vizsgán alkalmazott értékelési szempontrendszer alacsony inter-rater reliabilitást mutat, vagyis az értékelők közötti egyetértés mértéke mérsékelt lesz.
(2) Az alternatív, analitikus értékelési szempontrendszer alkalmazása növeli az értékelők közötti egyetértést és az értékelés következetességét.
(3) A pedagógusok által adott pontszámok jelentős eltéréseket mutatnak, különösen az érettségi vizsgán alkalmazott szempontok esetében, ami az értékelési rendszer szubjektivitását és következetlenségét jelzi.
(4) Az összbenyomás-alapú értékelés szoros összefüggést mutat az analitikus szempontokkal, de bizonyos tényezők kiemeltebb szerepet játszanak a pedagógusok végső döntésében.

A kutatás eredményei

A szempontsorok megbízhatósága

Az elemzés során sor került az értékelők közötti egyetértés (inter-rater reliabilitás) és a mérőeszközök belső konzisztenciájának a vizsgálatára, három megbízhatósági mutató (Cohen-féle kappa, Fleiss-féle kappa és McDonald's-féle ómega) összehasonlításával. A három mutató különböző aspektusok mentén teszi lehetővé az értékelési rendszer megbízhatóságának a vizsgálatát. A Cohen-féle kappa az értékelők páronkénti egyetértését méri, amely alkalmas az értékelési konzisztencia alapvető vizsgálatára. A Fleiss-féle kappa kiterjesztett változatként több bíráló esetén is alkalmazható, így pontosabb képet ad az összes értékelő közötti egyezésről. A McDonald's-féle ómega a belső konzisztencia vizsgálatára szolgál, amely meghatározza, hogy az egyes értékelési szempontok mennyire együttesen mérik ugyanazt a konstrukciót. A három mutató együttes alkalmazása biztosítja, hogy az értékelési rendszer megbízhatóságát mind az értékelők közötti egyezés, mind pedig a mérőeszköz struktúrájának a konzisztenciája szempontjából elemezni lehessen.

Az érettségi értékelési szempontrendszer megbízhatóságának vizsgálata során az elsődleges cél annak feltárása volt, hogy milyen mértékben mutatnak egyetértést a különböző pedagógusok ugyanazon szöveg értékelésekor. A statisztikai próbák eredményei azt mutatják, hogy az értékelők közötti egyetértés mindegyik vizsgált változó esetén (tartalom, szövegszerkezet, nyelvi igényesség, helyesírás, külalak, összpontszám) meglehetősen alacsony, amelyet a Cohen-féle kappa- és a Fleiss-féle kappa-mutatók is alátámasztanak.

A Cohen-féle kappa értéke az összes bevont változó tekintetében -0,011, amely mérsékelt egyetértési szintre utal (2. táblázat).

2. táblázat
A Cohen-féle kappa értéke az érettségi szempontsor alkalmazásakor

			95%-os konfidencia intervallum	
Értékelési szempontok	**Súlyozatlan kappa**	**Standard hiba (SE)**	**Alsó határ**	**Felső határ**
Kappa (átlag)	-0,011			
Szerkezet – nyelvi igényesség	-0,092	0,092	-0,271	0,088
Szerkezet – helyesírás (fordított)	0,0	0,0	0,0	0,0
Nyelvi igényesség – helyesírás (fordított)	0,0	0,0	0,0	0,0
Szerkezet – külalak	-0,003	0,025	-0,052	0,046
Nyelvi igényesség – külalak	-0,034	0,036	-0,103	0,036
Helyesírás (fordított) – külalak	0,0	0,0	0,0	0,0
Szerkezet – tartalom	0,0	0,0	0,0	0,0
Nyelvi igényesség – tartalom	0,0	0,0	0,0	0,0
Helyesírás (fordított) – tartalom	0,0	0,0	0,0	0,0
Külalak – tartalom	0,0	0,0	0,0	0,0
Szerkezet – összpontszám	0,0	0,0	0,0	0,0
Nyelvi igényesség – összpontszám	0,0	0,0	0,0	0,0
Helyesírás (fordított) – összpontszám	0,0	0,0	0,0	0,0
Külalak – összpontszám	0,0	0,0	0,0	0,0
Tartalom – összpontszám	-0,043	0,011	-0,064	-0,022

A különböző értékelési szempontok közötti egyetértés szintén gyenge, a legtöbb kappa-érték zéró körül mozog. Például a szerkezet és a nyelvi igényesség közötti kappa-érték -0,092, ami gyenge egyetértést jelez, míg más dimenziópárok kappa-értékei, például a tartalom és az összpontszám mutatói között szintén alacsonyak (-0,043). Ezek az értékek megerősítik azt a feltevést, hogy az értékelők eltérő módon ítélik meg a szempontokat.

A Fleiss-féle kappa értéke -0,061, amely szintén gyenge egyetértésre utal az értékelők között, míg a konfidenciaintervallum (-0,084 és -0,038) azt mutatja, hogy az értékelők közötti eltérések nagyobb mértékűek, mint ami egy következetes, standardizált értékelési rendszer esetében ideális lenne.

A McDonald's-féle ómega-mutató értéke 0,47, amely alacsony mértékű belső konzisztenciát jelez. Ez azt mutatja, hogy az egyes értékelési szempontok nem alkotnak egy jól definiált faktorstruktúrát, vagyis az értékelők nem egyforma mértékben veszik figyelembe a különböző kritériumokat. Ennek következtében az érettségi értékelési rendszer nem biztosít teljesen egységes alapot az értékeléshez, és az értékelők saját interpretációjuk szerint súlyozhatják az egyes szempontokat, ami jelentős variabilitáshoz vezet. Mindezek alapján megállapítható, hogy az érettségi értékelési szempontrendszere mérsékelt megbízhatóságot mutat, hiszen a kapott eredmények alapján az értékelők közötti egyetértés alacsony, a pontozás pedig következetlen a bírálók között.

Az alternatív értékelési szempontrendszer megbízhatóságának vizsgálata során ugyanezen mutatók elemzésére került sor. Az eredmények azt mutatják, hogy az új értékelési keretrendszer megbízhatóbbnak tekinthető, mint az érettségi vizsgán alkalmazott rendszer: a Cohen-féle kappa-érték az új szempontok esetében 0,199, amely ugyan szintén mérsékelt egyetértést jelez, de az értékelők ebben a rendszerben nagyobb fokú összhangot mutatnak a szöveg minősítésében (3. táblázat).

3. táblázat
A Cohen-féle kappa értéke az alternatív szempontsor alkalmazásakor

			95%-os konfidencia intervallum	
Értékelési szempontok	**Súlyozatlan kappa**	**Standard hiba (SE)**	**Alsó határ**	**Felső határ**
Kappa-érték átlaga	0,199			
Tartalom – feladattartás (szövegtípus)	0,494	0,112	0,275	0,712
Tartalom – feladattartás (hangnem)	0,444	0,110	0,228	0,661
Feladattartás (szövegtípus) – feladattartás (hangnem)	0,950	0,048	0,856	1,000
Tartalom – szerkezet	0,336	0,117	0,106	0,566
Feladattartás (szövegtípus) – szerkezet	0,713	0,109	0,499	0,927
Feladattartás (hangnem) – szerkezet	0,657	0,114	0,433	0,881
Tartalom – stílus	0,208	0,129	-0,044	0,461
Feladattartás (szövegtípus) – stílus	0,162	0,109	-0,051	0,375
Feladattartás (hangnem) – stílus	0,106	0,102	-0,093	0,306
Szerkezet és kidolgozás – stílus	0,075	0,102	-0,125	0,276
Tartalom – érthetőség	0,286	0,114	0,063	0,509
Feladattartás (szövegtípus) – érthetőség	0,656	0,115	0,431	0,882
Feladattartás (hangnem) – érthetőség	0,698	0,113	0,476	0,920
Szerkezet és kidolgozás – érthetőség	0,560	0,126	0,313	0,808
Stílus – érthetőség	0,020	0,094	-0,164	0,204
Tartalom – nyelvhelyesség	0,250	0,121	0,013	0,487
Feladattartás (szövegtípus) – nyelvhelyesség	-0,018	0,089	-0,192	0,156
Feladattartás (hangnem) – nyelvhelyesség	-0,074	0,076	-0,224	0,075
Szerkezet és kidolgozás – nyelvhelyesség	-0,087	0,081	-0,246	0,071
Stílus – nyelvhelyesség	0,268	0,141	-0,009	0,545
Érthetőség – nyelvhelyesség	-0,142	0,066	-0,272	-0,012
Tartalom – helyesírás	0,064	0,061	-0,056	0,184
Feladattartás (szövegtípus) – helyesírás	-0,026	0,025	-0,074	0,023
Feladattartás (hangnem) – helyesírás	-0,024	0,023	-0,069	0,021
Szerkezet és kidolgozás – helyesírás	-0,026	0,025	-0,074	0,023
Stílus – helyesírás	-0,041	0,039	-0,118	0,036
Érthetőség – helyesírás	-0,024	0,023	-0,070	0,021
Nyelvhelyesség – helyesírás	0,094	0,089	-0,080	0,268
Tartalom – külső megjelenés	0,268	0,103	0,067	0,470
Feladattartás (szövegtípus) – külső megjelenés	0,569	0,119	0,335	0,802
Feladattartás (hangnem) – külső megjelenés	0,609	0,118	0,377	0,841
Szerkezet és kidolgozás – külső megjelenés	0,563	0,123	0,321	0,804
Stílus – külső megjelenés	-0,019	0,077	-0,170	0,133
Érthetőség – külső megjelenés	0,504	0,133	0,242	0,756
Nyelvhelyesség – külső megjelenés	-0,079	0,061	-0,200	0,041
Helyesírás – külső megjelenés	-0,025	0,024	-0,071	0,022

Az egyes értékelési szempontok között néhány páronkénti kappa-érték különösen magasnak mutatkozik. Például a *feladattartás (szövegtípus)* és a *feladattartás (hangnem)* közötti kappa-érték 0,950, amely szinte tökéletes egyetértést jelent az értékelők között. Szintén magas egyetértést mutattak a *feladattartás (szövegtípus)* és a *szerkezet és kidolgozás* (0,713), valamint a *feladattartás (hangnem)* és a *szerkezet és kidolgozás* (0,657) közötti párosítások. Ez arra utal, hogy az értékelők ezekben a szempontokban egységesebben alkalmazták a kritériumokat, és az alternatív rendszer biztosította számukra a következetesebb pontozási lehetőséget.

A Fleiss-féle kappa-érték az alternatív szempontsor alkalmazásakor 0,147, amely szintén mérsékelt egyetértést jelez az értékelők között. A 95%-os konfidenciaintervallum (0,117 és 0,177) ugyanakkor azt mutatja, hogy az értékelők közötti összhang jóval stabilabb, mint az érettségi szempontsor esetében. A McDonald's-féle ómega-mutató az alternatív szempontrendszernél 0,91, amely magas belső konzisztenciára utal. Ez azt jelzi, hogy az egyes szempontok erősebb összefüggést mutatnak egymással, és az értékelési struktúra következetesebb, mint az érettségi vizsgán alkalmazott rendszerben.

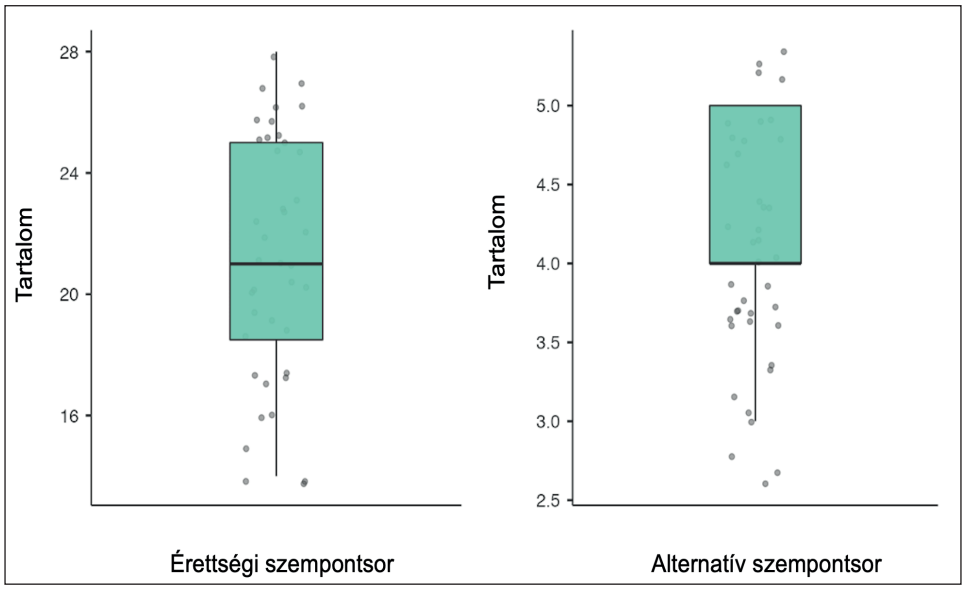
A két szempontrendszer összehasonlító elemzése egyértelműen azt mutatja, hogy az alternatív értékelési szempontrendszer megbízhatóbban működik, legalábbis a vizsgált minta alapján. Az érettségi szempontrendszer esetében a Cohen-féle kappa negatív vagy zéró közeli értékeket mutatott, míg az alternatív rendszerben erősebb egyetértés volt tapasztalható. A Fleiss-féle kappa-értékek szintén alacsonyabbak voltak az érettségi rendszerben, míg az új szempontrendszernél az egyetértési szintek magasabbak lettek. A McDonald's-ómega-mutató alapján az érettségi rendszer gyenge belső konzisztenciával rendelkezik (0,47), míg az alternatív rendszerben magasabb belső konzisztencia figyelhető meg (0,91). Ez azt jelenti, hogy az új értékelési rendszer struktúrája stabilabb, a pedagógusok között nagyobb egyetértés és belső konzisztencia figyelhető meg, és ez megbízhatóbb alapot ad a következetes és objektívabb értékelés megvalósítására.

A bíráló pedagógusok közötti különbségek

Az elemzés következő lépése annak megállapítása volt, hogy milyen mértékben mutatkoznak különbségek az egyes pedagógusok által adott pontszámok között. A fogalmazás értékelése akkor tekinthető következetesnek és megbízhatónak, ha a pedagógusok hasonló módon alkalmazzák az értékelési szempontokat, és pontozásukban nincs jelentős szórás vagy kiugró eltérés. Ennek vizsgálata szintén leíró statisztikai és grafikus elemzéseken, valamint Grubbs-teszten alapult, amelynek célja a kiugró értékek azonosítása volt az egyes változók esetében. Az adatsorok normál eloszlásának meghatározása Shapiro–Wilk-teszttel történt, a Grubbs-teszt alapját pedig a változók Z-score értékei képezték. Negyvenfős minta és 95%-os konfidenciaintervallum mellett 2,866 pontos Z-score abszolútérték felett beszélhetünk kiugró eltérésről.

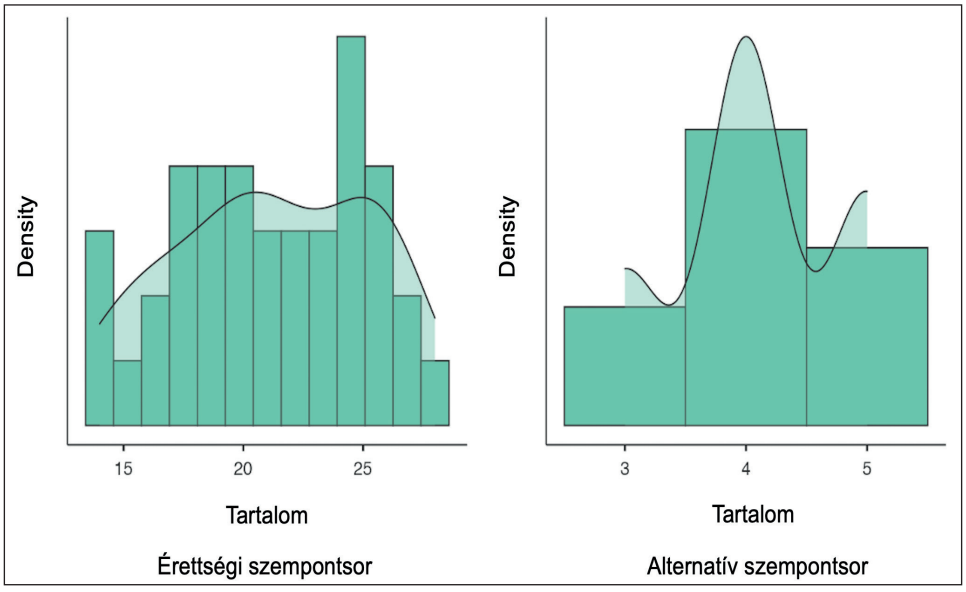
Az érettségi szempontrendszer alkalmazása során az értékelők pontszámai nagyobb szóródást mutattak, míg az alternatív rendszer esetében az értékelések egyértelműbben csoportosultak egy szűkebb tartományon belül. Az érettségi rendszer több szempontja esetében bimodális eloszlás volt megfigyelhető, vagyis a pedagógusok egy része jellemzően alacsonyabb, míg másik részük magasabb pontszámokat adott ugyanarra a szövegproduktumra. Ez különösen a tartalom és a szerkezet kategóriáiban volt észlelhető, ami arra utal, hogy az értékelők eltérően súlyozták az adott szempontokat.

A tartalom értékelésekor az érettségi rendszerben az átlagos pontszám 21,175 (a maximum 30-ból), a szórás 4,075, míg az alternatív rendszerben az átlag 4,1 (a maximum 5-ből), a szórás pedig 0,709 volt. Ez arra utal, hogy az érettségi értékelési rendszerben a pedagógusok nagyobb varianciát mutattak, míg az alternatív rendszerben a pontozás kisebb eltéréseket mutatott az egyes értékelők között (1. ábra).



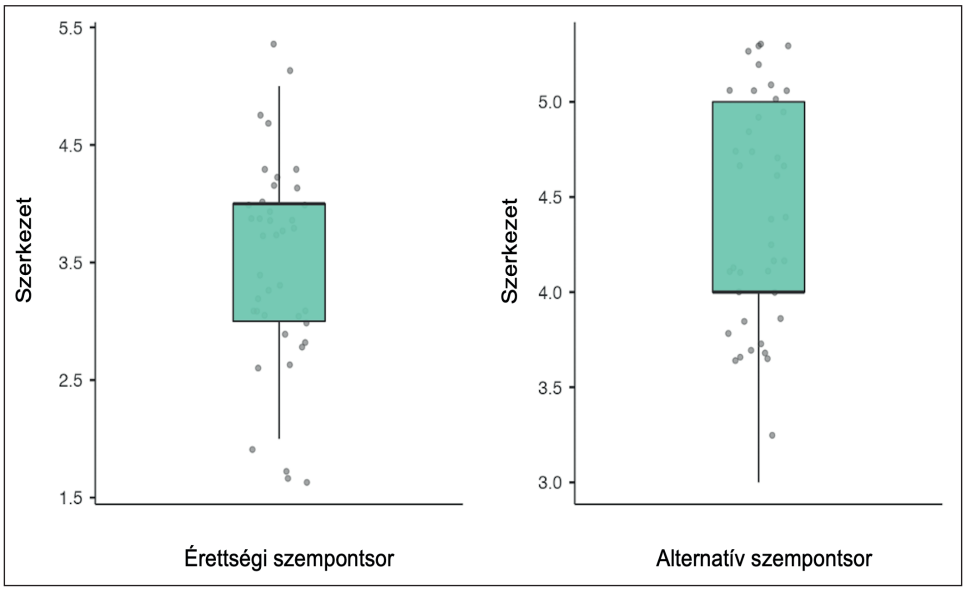
1. ábra
A tartalom pontszámai a két szempontsor alkalmazásával

A Shapiro–Wilk-teszt az érettségi rendszer esetében p = 0,071, ami a normális eloszlás határán van, de nem mutat extrém eltérést, míg az alternatív rendszerben p < 0,001, ami szignifikáns eltérést mutat a normál eloszláshoz képest, és azt jelzi, hogy a pedagógusok szorosabban csoportosultak egy adott értékelési sávon belül. A Grubbs-teszt alapján a Z-score értékek abszolútértékének maximuma az alternatív rendszerben 1,552, nem mutat kiugró értékeket, így az értékelések egyenletesek. A grafikus elemzés alapján az alternatív szempontsor esetén mutatkozik ugyan eltérés, ugyanakkor a leggyakoribb értékek a skála felső tartományában helyezkednek el, a szóródás pedig jellemzően az alsó tartományok irányába húz. Az érettségi rendszer esetén a grafikus elemzés rámutat arra, hogy az adatsor jellemzően bimodális természetű, a 18–20 és 24–26 közötti pontszámok a legjellemzőbbek, a leggyakoribb értékektől pedig mind a felső, mind pedig az alsó tartományokba viszonylag egyenletes a megoszlás (2. ábra).



2. ábra
A tartalomra kapott pontszámok megoszlása

A szerkezet értékelése szintén eltéréseket mutatott. Az érettségi rendszerben az átlag 3,525, a szórás 0,816, míg az alternatív rendszerben az átlag 4,45, a szórás 0,552 volt. Az érettségi szempontsorának alkalmazása során a pontozás szélesebb tartományban mozgott, míg az alternatív rendszerben az értékelések közelebb voltak egymáshoz, ami egyértelműbb kritériumrendszert és következetesebb értékelést jelez (3. ábra) Figyelemre méltó, hogy a maximálisan elérhető pontszám a szerkezet tekintetében mindkét szempontsor esetén 5 pont. Az eredményekben tapasztalt eltérést okozhatja az, hogy az érettségi értékelési rendszerben a szerkezet értékelése gyakran összemosódik más szempontokkal, leginkább a tartalommal, így az értékelők eltérő súlyozási stratégiákat alkalmazhatnak. Az alternatív szempontsor ezzel szemben igyekszik világosan elkülöníteni a szerkezeti elemek értékelését, egyértelműbb kritériumokat biztosítva ezáltal a bírálók számára és részletesebb iránymutatást adva arról, hogy milyen szövegjellemzők indokolják a különböző pontszámokat.

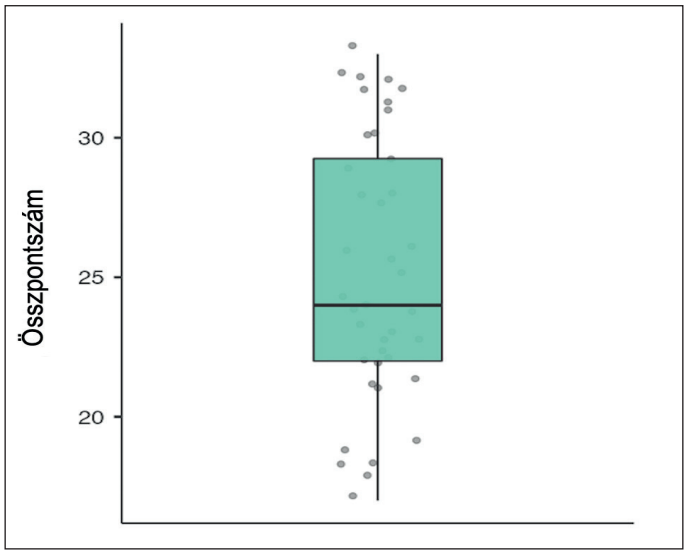


3. ábra
A szerkezet pontszámainak a megoszlása

Az érettségi szempontsor további szempontjaira adott pontszámok között szintén mutatkoznak eltérések. A nyelvi igényesség esetében az átlagos pontszám 3,225, a medián 3, a szórás pedig 0,862, ami mérsékelt szóródást jelez. A Shapiro–Wilk-teszt eredménye p < 001, vagyis az eloszlás szignifikánsan eltér a normálistól, ami arra utal, hogy az értékelők eltérő módon pontozták ezt a szempontot is. Ugyanakkor a Grubbs-teszt legnagyobb Z-score értéke 2,059, amely nem haladja meg a kritikus küszöbértéket, vagyis nem azonosítható szélsőséges eltérés egyes értékelők pontszámai között. A grafikus elemzés alapján az értékelések itt is bimodális eloszlást mutattak, a pontszámok 3 és 4 között koncentrálódtak, ez arra utal, hogy az értékelők két eltérő értelmezési módot alkalmaztak az értékelés során.

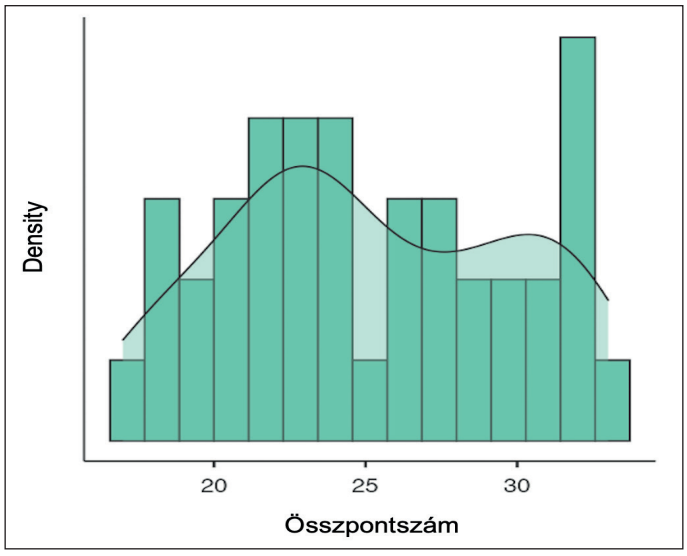
A helyesírás értékelésében, amely az érettségin fordított skálázású, az átlag -4,35, a szórás 0,43, amely kis szóródást mutat, tehát a pedagógusok viszonylag egységesen pontozták ezt a szempontot az érettségi útmutatója alapján. A külalak szempontjában az átlagos pontszám az érettségi szempontrendszer eredményeiben 1,525, a medián 1,5, a szórás pedig 0,506, ez rendkívül kis szóródást mutat. Ez azt jelzi, hogy az értékelők szinte teljesen egyöntetűen pontozták ezt a kategóriát. Az alacsony szórás magas fokú következetességet jelez, ugyanakkor ne felejtsük el, hogy ez a szempont az érettségi esetén maximum két pontot érhet.

Az összpontszámok tekintetében az értékelők átlagos pontszáma 25,2, a medián 24, a szórás 4,746, ez mérsékelt szóródásra utal, tehát a pedagógusok pontozása viszonylag következetes volt, noha az eltérések nem elhanyagolhatók (4. ábra).



4. ábra
Az összpontszám megoszlása az érettségi szempontsorával történő javításkor

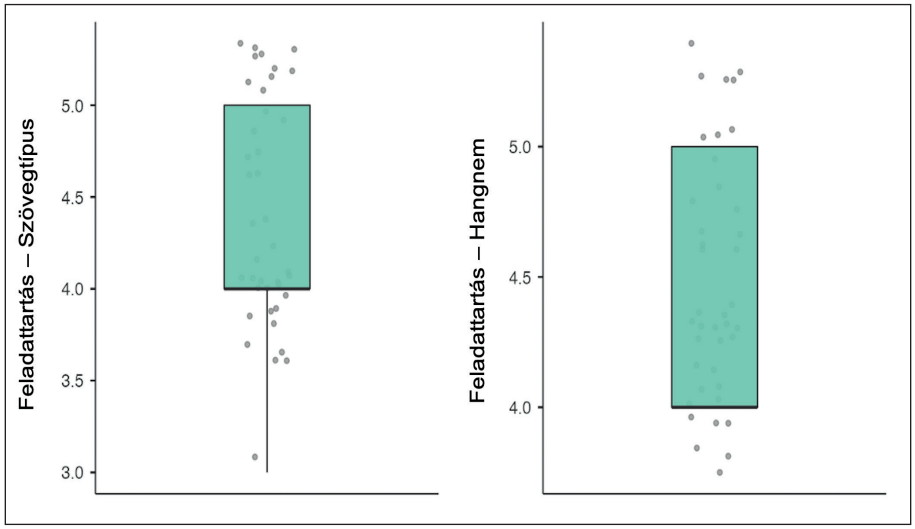
A Shapiro–Wilk-teszt alapján: p = 0,041, vagyis az összpontszámok eloszlása statisztikailag szignifikánsan eltér a normálistól, amit a grafikai elemzés is megerősített. Az eloszlás ezek alapján itt is bimodális jellegű, a leggyakoribb pontszámok 22–24 és 30–32 között csoportosultak, ez megerősíti a tartalomnál látott eredményekből következő azon feltevést, hogy a javítók két eltérő megközelítést alkalmaztak a fogalmazás értékelése során (5. ábra). A Grubbs-teszt eredményei szerint azonban nem voltak szélsőségesen kiugró pontszámok, mivel a maximális Z-score érték 1,728, amely nem haladja meg a kritikus küszöbértéket.



5. ábra
Az érettségi útmutató alapján értékelt szöveg összpontszámának megoszlása

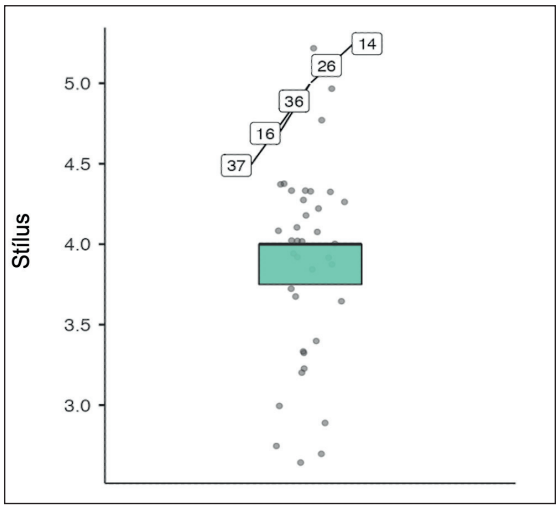
Az alternatív szempontrendszer alkalmazása során az értékelők pontszámai összességében kisebb szóródást mutattak, és az értékelési eredmények is következetesebbek voltak az érettségi rendszerhez képest. Az egyes értékelési kategóriákban ugyan előfordultak eltérések, ezek mértéke azonban kevésbé volt jelentős, és a kiugró értékek száma is alacsony maradt. A *feladattartás (szövegtípus)* esetében az átlagos pontszám 4,4, a szórás 0,545, amely viszonylag egyenletes értékelési mintázatra utal. A Shapiro–Wilk-teszt szignifikáns eltérést mutatott a normálishoz képest, a Grubbs-teszt alapján azonban nem voltak szélsőséges eltérések. A grafikus elemzés szerint az értékelések itt is bimodális eloszlást mutattak, a pontszámok többsége 3 és

4 között koncentrálódott, ami azt jelzi, hogy az értékelők között volt némi eltérés ebben a szempontban, de a különbségek nem voltak kritikusak. A *feladattartás (hangnem)* esetében hasonló eredményeket találunk, az átlag 4,425, a szórás 0,501, amely alacsony varianciát jelez. Az értékelők ezen a területen is egyetértést mutattak, amit a Grubbs-teszt eredménye is alátámasztott, hiszen szélsőséges eltérés nem volt kimutatható. Az értékelések többsége a 4 és 5 pont közötti tartományban helyezkedett el, így az eltérések minimálisak maradtak (6. ábra).



6. ábra
A feladattartásra kapott pontszámok megoszlása az alternatív szempontsor alkalmazásakor

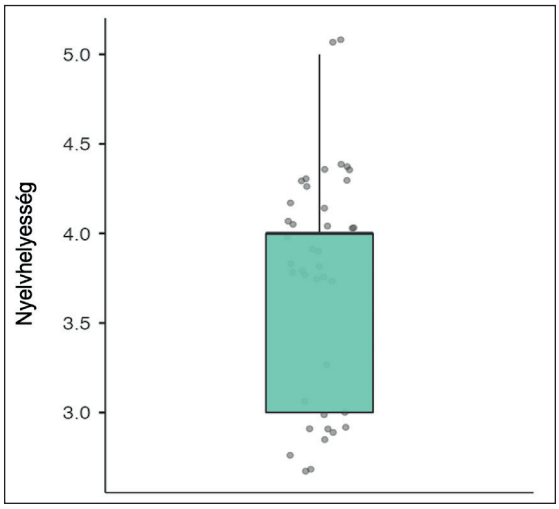
A stílus értékelése során az átlagos pontszám 3,875, a szórás 0,607, amely kissé nagyobb szórást jelez, mint az előző szempontoknál. A Shapiro–Wilk-teszt ismételten szignifikáns eltérést mutatott a normálishoz képest, a Grubbs-teszt alapján azonban az eltérések itt sem tekinthetők kritikusnak. Az adatok elemzése szerint a 4 pont volt a leggyakoribb érték, de öt pedagógus adott maximális pontszámot, ami enyhe kiugrásként jelenhetett meg az eloszlásban (7. ábra).



7. ábra
A stílusra kapott pontszámok megoszlása az alternatív szempontsor alkalmazásakor

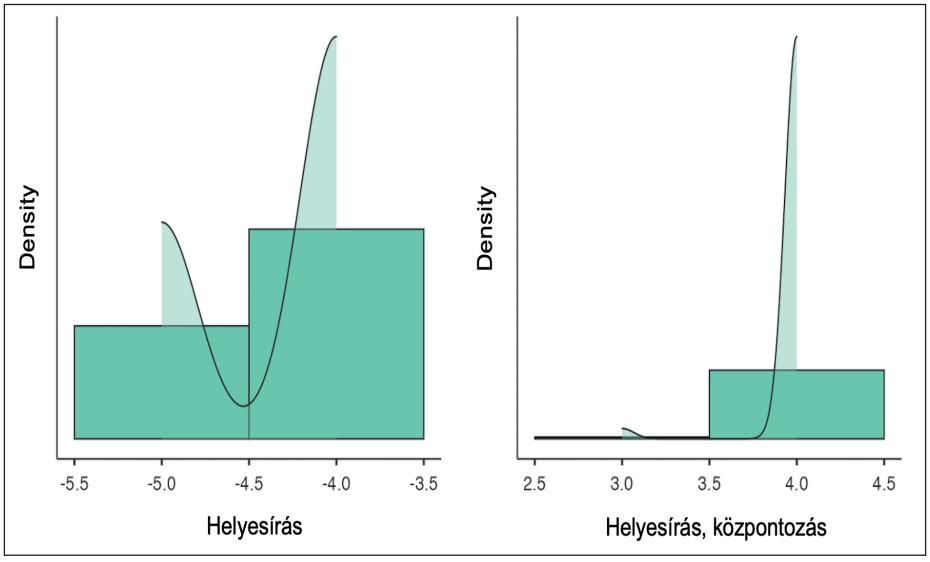
Az érthetőség kategóriában az átlag 4,475, a szórás 0,506, ez tovább erősíti azt a tendenciát, hogy az értékelések ezen szempontok mentén kevésbé szóródtak. Az értékelések többsége a 4 és 5 pont között helyezkedett el. A nyelvhelyesség esetében az átlagos pontszám 3,75, a szórás 0,543, amely közepes

szóródást mutat. Az értékelések főként a 3 és 4 pont között mozogtak, tehát itt tapasztalható volt némi eltérés az értékelők között (8. ábra).



8. ábra
A stílusra kapott pontszámok megoszlása az alternatív szempontsor alkalmazásakor

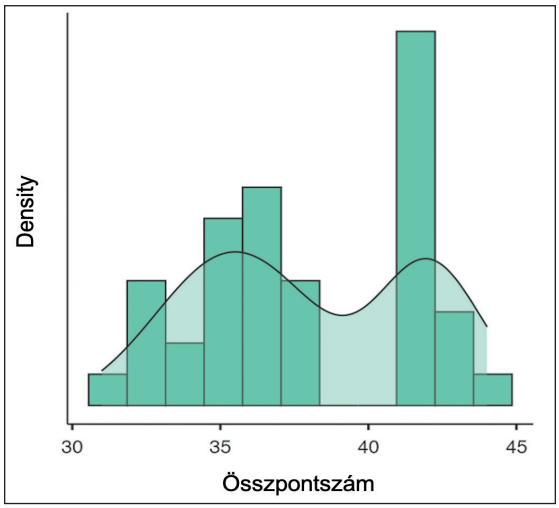
A helyesírás átlagos pontszáma 3,975, a szórás 0,158, és a Shapiro–Wilk p-értéke $p < 0,001$, ami szignifikáns eltérést mutat a normál eloszláshoz képet. A Grubbs-teszt Z-score abszolútérték maximuma 6,166, amely jelentős kiugrást jelez. A grafikus elemzés ugyanakkor rámutat arra, hogy a 4 pont a legjellemzőbb, egyetlen esetben született 3-as osztályzat, amely előidézte a statisztikailag is jelentős kiugrást. Amennyiben ezt az értéket nem vesszük figyelembe, az osztályzatok konzisztens formában 4 pontot mutatnak, ez kifejezetten erős egyetértést jelez a helyesírás szempontjának pontozásában az értékelők között, és jóval nagyobb hasonlóságot, mint az érettségi szempontsor helyesírás pontszámaiban (9. ábra).



9. ábra
A helyesírásra adott pontszámok eloszlása az érettségi és az alternatív szempontsor alkalmazásakor

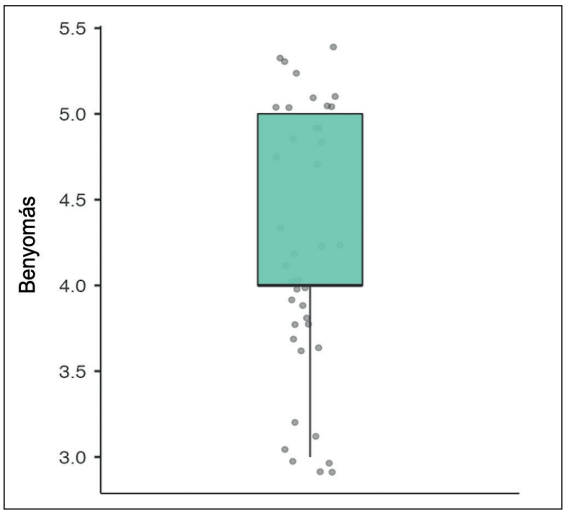
A külalak kategóriában az átlag 4,575, a szórás 0,501, amely minimális eltérést mutat az értékelők között. Az adatok szerint az 5 pont volt a leggyakoribb érték, amelyet a 4-es osztályzat követett, más pontszámot nem adtak az értékelők. Ez arra utal, hogy ebben a szempontban az értékelők szinte teljesen egyöntetűek voltak, akárcsak az érettségi szempontsor alkalmazásával pontozott külalak esetében.

Az összpontszámok tekintetében az átlag 38,025, a szórás 3,683, ami alacsonyabb varianciát mutat az érettségi értékelésekhez képest. A Shapiro–Wilk-teszt szignifikáns eltérést mutatott a normál eloszláshoz képest ($p = 0,006$), de a Grubbs-teszt alapján nem volt szélsőséges eltérés az értékelések között. Az értékelések bimodális jellegűek voltak, a leggyakoribb pontszámok 35 és 41 között mozogtak (10. ábra).



10. ábra
Az alternatív szempontsor alapján értékelt szöveg összpontszámának a megoszlása

Az összbenyomás értékelésénél az átlag 4,225, a szórás 0,733, ez mérsékelt szóródást jelez. A Shapiro–Wilk-teszt szerint az eloszlás szignifikánsan eltért a normálistól, a Grubbs-teszt alapján azonban nem volt azonosítható szignifikáns kiugrás. A leggyakoribb pontszámok 35 és 41 között helyezkedtek el, hasonlóan az összpontszámokhoz (11. ábra).



11. ábra
Az összbenyomásra kapott pontszámok megoszlása az alternatív szempontsor alkalmazásakor

Összességében a pedagógusok értékelései az alternatív szempontrendszerben következetesebbek voltak, ugyanakkor egy-egy szempont esetében felmerültek enyhe mértékű eltérések. Bár a Shapiro–Wilk-teszt minden szempontban szignifikáns eltérést mutatott a normál eloszláshoz képest, a Grubbs-teszt nem jelzett számottevő kiugró értékeket. A pontszámok jellemzően a felsőbb tartományokban mozogtak, ami magyarázatot ad a normál eloszlás hiányára, ugyanakkor szignifikáns különbségek nem állapíthatók meg köztük.

Ahhoz, hogy választ kapjunk arra a kérdésre, hogy a két szempontrendszer közül melyikben mutatkozik meg kisebb különbség a bíráló pedagógusok pontszámai között, érdemes táblázatba foglalni a Z-score értékek minimum- és maximumértékét a szempontok, ezek átlaga és a végső összpontszám szerint (4. táblázat).

4. táblázat
A szempontokra kapott Z-score értékek minimum- és maximumértékei a két mérőeszköz szempontjaiban

Érettségi szempontrendszer, Z-score			Alternatív szempontrendszer, Z-score		
	Minimum	**Maximum**		**Minimum**	**Maximum**
Tartalom	-1,76	1,67	Tartalom	-1,55	1,27
Szerkezet	-1,87	1,81	Szövegtípus	-2,57	1,10
Ny. igényesség	-1,42	2,06	Hangnem	-0,85	1,15
Helyesírás	-1,35	0,72	Szerkezet	-2,62	1,00
Külalak	-1,04	0,94	Stílus	-1,44	1,85
Átlag	**-1,488**	**1,44**	Érthetőség	-0,94	1,04
Összesen	**-1,73**	**1,64**	Nyelvhelyesség	-1,38	2,30
			Helyesírás	-6,17	0,16
			Külalak	-1,15	0,85
			Átlag	**-2,07**	**1,19**
			Összesen	**-1,91**	**1,62**
			Benyomás	-1,67	1,06

A táblázat az érettségi és az alternatív szempontrendszer szerint kapott Z-score értékek minimum- és maximumhatárait mutatja be az egyes értékelési szempontok mentén. A Z-score értékek az átlagos eltéréseket jelölik az értékelői pontszámok esetében, így minél kisebb tartományon belül mozognak, annál egységesebb az értékelés. Az érettségi rendszerben az értékelők közötti eltérés kisebb tartományban mozog, az átlagos minimum -1,488, a maximum pedig 1,44, míg az összpontszámok minimuma -1,73, maximuma pedig 1,64. Ez arra utal, hogy az értékelők pontozása kevésbé tért el az átlagtól, ami konzervatívabb, de potenciálisan kevésbé differenciáló értékelési gyakorlatra utal. Az alternatív szempontrendszerben az átlagos minimum -2,07, a maximum 1,19, az összpontszámok minimuma -1,91, maximuma pedig 1,62. Ez azt sugallja, hogy az értékelések nagyobb szórást mutattak az alsó tartományban, a felső tartományban azonban kisebb eltérések figyelhetők meg, ami kiegyensúlyozottabb értékelési folyamatra utal. Az összesített pontszámok esetében az érettségi rendszer -1,73 és 1,64 közötti, míg az alternatív rendszer összességében hasonló, de kissé szűkebb felső tartományú szórást eredményezett.

Az érettségi szempontrendszert használva a bírálók között a legnagyobb különbség a szerkezet szempontjában figyelhető meg, ezt követi a nyelvi igényesség, majd a tartalom. Az alternatív értékelési szempontsor adatait tekintve megállapítható, hogy a bírálók közötti eltérések általában 3,0 körüli terjedelműek, azaz az értékelések viszonylag következetesek voltak.

Az alternatív szempontrendszer esetén a legkisebb eltérések az érthetőség (1,98) és a külalak (2,00) szempontokban figyelhetők meg, ez arra utal, hogy ezeket a kategóriákat az értékelők szinte azonos módon pontozták. A legnagyobb eltérések a helyesírás (6,33), a nyelvhelyesség (3,68), a szövegtípus (3,67) és a szerkezet (3,62) kategóriákban figyelhetők meg. Ezek közül kiemelkedik a helyesírás, ahol a minimumérték -6,17, a maximum pedig 0,16, ami jelentősen nagyobb szórást mutat, mint a többi szempont. Ez az eltérés azonban nem általános jelenséget tükröz, hanem egyetlen kiugró értékelés hatását: a 40 bírálóból 39-en egyöntetűen 4-es pontszámot adtak, míg egyetlen értékelő 3-asra ítélte a helyesírást, ami torzította a statisztikai szóródást. Ez jól mutatja, hogy a Grubbs-teszt által kimutatott extrémértékek mögött olykor egy-egy eltérő értékelési döntés

állhat, amely nem az egész rendszer működését jellemzi. A szempontok átlagos eltérésének terjedelme 3,26, míg az összpontszámok esetében 3,53, ami azt mutatja, hogy a pedagógusok közötti különbségek viszonylag mérsékeltek voltak az értékelési szempontok összességét tekintve. Az összbenyomás értékelése szintén egységes volt (terjedelme: 2,73), ez arra utal, hogy az értékelők globális ítéletei kevésbé tértek el egymástól, és az értékelési folyamat során hasonló végső benyomást alakítottak ki a tanulói teljesítményről.

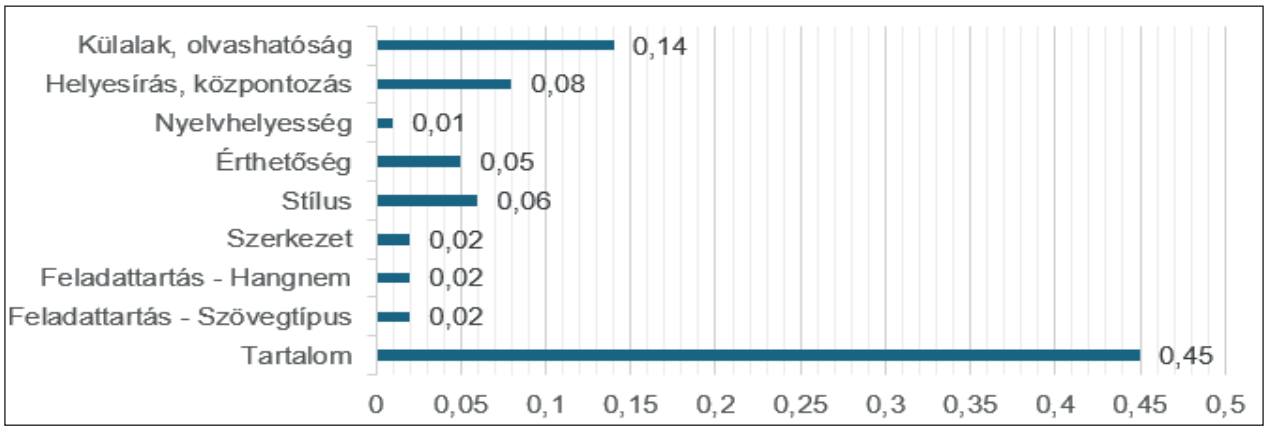
Az eredmények összegzéseként megállapítható, hogy az alternatív szempontrendszer alkalmazása csökkentette az értékelők közötti eltéréseket, és ez az összpontszámokban is megmutatkozik. Bár az érettségi rendszerben kisebb szórás volt tapasztalható, ez nem feltétlenül az értékelés következetességét, hanem a pontozási sávok kevésbé differenciáló jellegét tükrözi. Az alternatív rendszer egyértelműbb kritériumai és az egységesebb szempontrendszer hozzájárultak a bírálói döntések konzisztenciájához (például a szerkezet szempontja), miközben lehetővé tették a teljesítmények pontosabb megkülönböztetését is.

Az analitikus szempontok és a globális ítélet kapcsolata az alternatív mérőeszköz eredményeiben

Az értékelési folyamat kulcskérdése az is, hogy a pedagógusok által adott összbenyomás mennyiben támaszkodik az objektíven mérhető analitikus szempontokra, vagyis mennyire szisztematikus a tanárok végső ítélete, és milyen tényezők befolyásolják ezt a legnagyobb mértékben. E célból a vizsgálat generalizált lineáris modellt alkalmazott, amelyben az összbenyomás szolgált függő változóként, míg az értékelési szempontok független változóként szerepeltek. A modell célja az volt, hogy megmutassa, a pedagógusok összbenyomása mennyire támaszkodik az analitikus szempontokra, és mely tényezők gyakorolják a legnagyobb hatást a végső értékelésre.

Az elemzés eredményei azt mutatják, hogy az összbenyomás rendkívül szorosan összefügg az egyes analitikus értékelési szempontokkal, hiszen a modell R-négyzet-értéke 0,93, ami azt jelenti, hogy a kilenc szempont az összbenyomás varianciájának 93%-át magyarázza. Ez arra utal, hogy az értékelők összegző ítélete nem véletlenszerű vagy szubjektív, hanem nagymértékben az egyes részszempontok alapján formálódik.

A szempontok hatása ugyanakkor nem egyenlő mértékű. Az eredmények szerint a tartalom a legerősebb hatású tényező, amely szignifikánsan és dominánsan befolyásolja az összbenyomást ($\eta^2p = 0,45$, $p < 0,001$). Ez azt jelzi, hogy a pedagógusok végső értékelésének középpontjában a szöveg tartalmi kidolgozottsága és gondolati mélysége áll, és ennek a szerepe messze meghaladja a többi szempontét (12. ábra).



12. ábra
Az analitikus szempontok határméretértéke (η^2p)

Az eredmények azt mutatják, hogy az értékelők az összbenyomás kialakításakor és pontozásakor leginkább a tartalomra fókuszálnak, ezt tekintik a legfontosabb minőségi tényezőnek. A külalak szintén hangsúlyosan befolyásolja az összbenyomást, valószínűleg azért, mert az esztétikailag rendezett, átlátható szövegek

könnyebben befogadhatók és olvashatók, ez indirekt módon pozitív hatással lehet az értékelői megítélésre. Az, hogy a többi szempont nem mutatott szignifikáns hatást a végső benyomásra, arra utalhat, hogy ezek a szempontok nem önálló értékelési tényezőként jelennek meg a pedagógusok gondolkodásában, hanem sokkal inkább a tartalomhoz kapcsolódva. Egy világosan felépített, gördülékenyen megfogalmazott szöveg természetesen jobb összbenyomást kelt, de az értékelők ezt valószínűleg már eleve a tartalom értékelésének a részeként kezelik.

Ezt megerősíti az alternatív értékelési szempontsor eredményeinek korrelációs elemzése is. Ebből kiderül, hogy a pedagógusok összbenyomása szoros összefüggésben áll az analitikus értékelési kritériumokkal, vagyis a szöveg átfogó megítélése nem elszigetelten, hanem az egyes szempontok együttes hatásaként alakul ki. A legerősebb kapcsolat a tartalommal (r = 0,88) és a feladattartás szempontjaival (r = 0,98) mutatkozik, ez azt jelzi, hogy az értékelő pedagógusok számára ezek a tényezők kiemelt szerepet játszanak a döntéshozatal során. A szerkezet, az érthetőség és a stílus szintén jelentős, de valamivel kisebb súlyú tényezők az összbenyomás meghatározásában, míg a helyesírás és a külalak gyengébb korrelációt mutatott, ez megerősíti a korábbi fogalmazásvizsgálatok eredményeit is (például Nagy 2013: 174). Mindez arra utal, hogy ezeket az értékelők inkább önállóan kezelték (5. táblázat).

5. táblázat
Az alternatív szempontsor eredményeinek szempontonkénti összefüggéseinek az erőssége

	T	Fsz	Fh	SzK	S	É	Ny	H	K
Tartalom									
Feladattartás (szövegtípus)	0,77								
Feladattartás (hangnem)	0,76	0,98							
Szerkezet és kidolgozás	0,67	0,72	0,70						
Stílus	0,51	0,53	0,51	0,63					
Érthetőség	0,65	0,70	0,70	0,60	0,62				
Nyelvhelyesség	0,60	0,51	0,49	0,65	0,45	0,64			
Helyesírás	0,24	0,13	0,14	0,14	0,04	0,15	0,23		
Külső megjelenés	0,63	0,64	0,64	0,62	0,59	0,51	0,45	0,19	
Összpontszám	0,87	0,84	0,83	0,85	0,75	0,82	0,75	0,24	0,76
Összbenyomás	0,88	0,90	0,89	0,77	0,65	0,76	0,61	0,25	0,76

Megjegyzés: minden összefüggés $p < 0,001$ szinten szignifikáns; T = tartalom; Fsz = feladattartás (szövegtípus); Fh = feladattartás (hangnem); SzK = szerkezet és kidolgozás; S = stílus; É = érthetőség; Ny = nyelvhelyesség; H = helyesírás; K = külső megjelenés

Az eredmények alapján tehát azt látjuk, hogy az alternatív szempontsor alkalmazása esetén az összbenyomás szorosan összefügg az analitikus értékelési szempontokkal, és az értékelők döntése nagy részben objektív tényezőkön alapul. Az összbenyomást elsősorban a tartalom minősége határozta meg. A többi szempont indirekt módon befolyásolja az értékelést, de amikor minden szempont együttesen szerepel a modellben, ezek hatása kevésbé kifejező. Mindez azt mutatja, hogy az alternatív értékelési rendszer jól tükrözi az értékelési folyamat súlypontjait, és az összbenyomás kialakulása szorosan követi az analitikus értékelési szempontok mentén felállított struktúrát.

Következtetések és összegzés

A kutatás célja az volt, hogy empirikus adatok alapján vizsgálja a 2024-től alkalmazott középszintű érettségi javítási-értékelési útmutatójának megbízhatóságát, összevetve egy alternatív, analitikusabb mérőeszköz alkalmazásával. A vizsgálat eredményei összességében azt mutatják, hogy az érettségi értékelési

rendszerében az értékelők közötti egyetértés alacsonyabb, az értékelési szempontok differenciáló ereje pedig mérsékelt, míg az alternatív szempontsor egységesebb, strukturáltabb értékelési gyakorlatot eredményezett.

A fejezet elején megfogalmazott hipotézisek vizsgálata a következő eredményeket mutatja: a statisztikai elemzések – a Cohen-féle és a Fleiss-féle kappa-értékek – alátámasztották, hogy az érettségi szempontsor alkalmazása során az értékelők közötti egyezés szintje alacsony. Az értékelési kritériumok szubjektív interpretációja és a szélesebb értékelési skála miatt az értékelők sokszor jelentősen eltérő pontszámokat adtak ugyanazon szövegproduktum egyes szempontjaira.

Az alternatív szempontsor alkalmazásával az értékelések közötti variancia megmaradt, ugyanakkor csökkent; az értékelők egységesebb módon pontozták a fogalmazást, és az értékelés megbízhatósági mutatói (Cohen-kappa, Fleiss-kappa, McDonald's ómega) is magasabb értékeket mutattak. Az alternatív szempontsor egyértelműbb kritériumai és részletesebb pontozási útmutatója segítette a bírálók döntéseinek az egységesítését, ez növelte a mérőeszköz objektivitását.

Az érettségi értékelési útmutató szerint végzett értékelések nagyobb szórást mutattak, és több esetben is bimodális eloszlás volt megfigyelhető, ami arra utal, hogy a pedagógusok eltérő módon súlyozták az egyes szempontokat. Ez különösen igaz volt a tartalom, a szerkezet és az összpontszám kategóriáiban. Az alternatív szempontsor esetén az értékelési eredmények szűkebb tartományban mozogtak, és a kiugró értékek száma is csökkent.

A generális lineáris modell elemzése alapján az összbenyomás leginkább a tartalom minőségétől függött, amelynek hatása kiemelkedően magas volt, míg a többi értékelési szempont hatása kevésbé volt szignifikáns, amikor a modellben együttesen jelentek meg. Ez azt jelzi, hogy az értékelők elsődlegesen a tartalmi szempontokat mérlegelték az értékelés során, míg a szerkezeti és a nyelvi igényesség tényezői indirekt módon hatottak a végső pontszámra.

A vizsgálat eredményeinek értelmezésekor ugyanakkor fontos figyelembe venni a kutatás néhány korlátját. Az alternatív mérőeszköz jelen formájában alkalmasnak mutatkozik az értelmező-érvelő jellegű szövegalkotási teljesítmények vizsgálatára, viszont a műértelmezés és az érvelés határterületén helyezkedik el. Ez a sajátosság részben magyarázhatja a tartalom dimenziójának kiemelt szerepét az értékelői döntésekben, de azt is jelzi, hogy a különböző szövegalkotási feladattípusok értékelése eltérő mértékben igényli a feladat jellegéhez igazított értékelési szempontokat. Az érettségi szövegalkotási feladat utasítása retorikai szituációt teremt, amelyben a vizsgázónak a megadott kérdésekre és szempontokra kell illeszkedő, tárgyszerű választ adnia. A műértelmező feladat esetében ez azt jelenti, hogy a tartalmi teljesítmény megítélése szorosan összefügg azzal, milyen mértékben és milyen minőségben reagál a tanuló a feladatban kijelölt értelmezési fókuszokra. A jelen kutatásban alkalmazott alternatív analitikus mérőeszköz ezzel szemben általánosabb szövegminőségi dimenziókra épült, és nem bontotta tovább a tartalmi teljesítményt a feladatban expliciten megfogalmazott részfeladatok mentén. Bár ez a megközelítés lehetővé tette az értékelők közötti eltérések empirikus összehasonlítását és a mérőeszközök megbízhatóságának a vizsgálatát, egy feladatspecifikusabb, a retorikai szituáció elemeit expliciten figyelembe vevő tartalmi szempontsor várhatóan tovább csökkentené az értékelés szubjektivitását. Ennek vizsgálata azonban már más módszertani keretek között, módosított, továbbfejlesztett mérőeszközzel lenne megvalósítható.

A kutatás eredményei mindezek ellenére rávilágítanak arra, hogy a 2024-től alkalmazott érettségi javítási-értékelési útmutatója továbbfejleszthető. Az értékelési szempontok operacionalizáltságának hiányosságai miatt a pedagógusok eltérő módon alkalmazzák a pontozási skálát, ami az értékelés szubjektivitását és következetlenségét is növeli. Az alternatív értékelési rendszer viszont strukturáltabb, megbízhatóbb keretet biztosított az értékelők számára; a különbségek továbbra is jelen vannak, viszont az analitikusabb szempontsor csökkentette ezeket, és növelte az értékelési folyamat objektivitását. A bemutatott alternatív szempontrendszer

sem tekinthető véglegesnek: további finomhangolást igényel az egyes értékelési dimenziók közötti átmenetek pontosítása, hiszen az analitikus megközelítés önmagában nem garancia az értékelés megbízhatóságára, amennyiben az értékelési kategóriák definíciói nem elég világosak, illetve ha az értékelői ítéletalkotás mögött nem áll stabil, közösen értelmezett kritériumrendszer. Az analitikus rendszer további fejlesztésének az iránya ezért az értékelési szintek közötti különbségek empirikus validálása, a kategóriák leíró pontosságának növelése, valamint a normatív mintaszövegek és illusztratív példák bevonása lehet.

Indokolt lehet az analitikus értékelési rendszer bevezetése vagy részleges integrálása a jelenlegi értékelési gyakorlatba. Az alternatív szempontsor alkalmazása a kutatás eredményei alapján egyértelműen csökkentette az értékelők közötti varianciát és strukturáltabb, átláthatóbb értékelési folyamatot eredményezett. Az analitikus megközelítés előnye, hogy egyértelmű teljesítményszinteket határoz meg több szempontban, ezáltal minimalizálja a pontozásban tapasztalható szubjektivitást. Az analitikus értékelési rendszer alkalmazása elősegíthetné a részletesebb visszacsatolást is, amely hozzájárulhatna a tanulók szövegalkotási készségének a fejlesztéséhez is.

A kutatás eredményei implikálják azt is, hogy az értékelési konzisztencia növeléséhez szükség van az értékelők képzésére is. A pedagógusok eltérő értékelési gyakorlata abból is eredhet, hogy az értékelési szempontok alkalmazásában nem rendelkeznek egységes referenciapontokkal. A fogalmazásértékeléssel kapcsolatos továbbképzések és értékelői műhelyek szervezése szintén segíthetné a következetesebb értékelési gyakorlat kialakítását.

Irodalom

Aslim Yetis, Veda 2019. Evaluating essay assessment: Teacher-developed criteria versus rubrics. Intra/inter reliability and teachers' opinions. *Croatian Journal of Education – Hrvatski časopis za odgoj i obrazovanje* 21(1): 103–129. https://doi.org/10.15516/cje.v21i1.2922

Eckes, Thomas 2008. Rater types in writing performance assessments: a classification approach to rater variability. *Language Testing* (25)2: 155–185. https://doi.org/10.1177/02655322070867

Fülöp Károly 2017. Esszészövegek szerkezetének kvantitatív elemzése. *Anyanyelv-pedagógia* (10)2: 34–47. https://doi.org/10.21030/anyp.2017.2.3

Horváth Zsuzsanna 1998. *Anyanyelvi tudástérkép. Középiskolai tantárgyi feladatbankok III.* Országos Közoktatási Intézet. Budapest.

Juhász Milán 2021. Az érvelő esszé értékelési szempontjainak megítélése pályakezdő és gyakorlott pedagógusok által. *Anyanyelv-pedagógia* (14)1: 23–39. https://doi.org/10.21030/anyp.2021.1.2

Juhász Milán 2022. A műértelmező szövegek értékelésének megbízhatósága a magyar nyelv és irodalom középszintű érettségi vizsgán. In: Molnár Dániel – Molnár Dóra – Nagy Adrián Szilárd (szerk.) *Tavaszi Szél 2022 / Spring Wind 2022 Tanulmánykötet II.* Doktoranduszok Országos Szövetsége (DOSZ). Budapest. 494–505.

Juhász Milán 2025. *A szövegalkotás mérésének módszertana.* Doktori disszertáció. ELTE BTK. Budapest.

Kádárné Fülöp Judit 1990. *Hogyan írnak a tizenévesek? Az IEA-fogalmazásvizsgálat Magyarországon.* Akadémiai Kiadó. Budapest.

Karkó Ádám 2022. Érettségi változások: magyar nyelv és irodalom. A közismereti érettségi vizsgatárgyak 2024. május-júniusi vizsgaidőszaktól érvényes vizsgakövetelményei. *Új Köznevelés* 78(1): 7–9.

Kerner Anna 2009. Esszé a középiskolában. A filozófiai esszé. *Anyanyelv-pedagógia* (2)2: 57–65. https://doi.org/10.21030/anyp.2009.2.6

Li, Wentao 2022. Scoring rubric reliability and internal validity in rater-mediated EFL writing assessment: Insights from many-facet Rasch measurement. *Reading and Writing* 35: 2409–2431. https://doi.org/10.1007/s11145-022-10279-1

Molnár Edit Katalin 2000. A fogalmazási képességek fejlődésének mérése. *Iskolakultúra* 10(8): 49–59.

Nagy Zsuzsanna 2009. 17 éves tanulók szövegalkotási képessége és szövegekre vonatkozó ítéletei. *Iskolakultúra* (19)11: 19–31. https://www.iskolakultura.hu/index.php/iskolakultura/article/view/20922 (2025. október 20.)

Nagy Zsuzsanna 2013. A fogalmazásértékelés megbízhatósága két független bíráló értékítéleteinek elemzése alapján. *Magyar Pedagógia* 3: 153–179.

Orosz Sándor 1972. *A fogalmazástechnika mérésmetodikai problémái és országos színvonala*. Tankönyvkiadó. Budapest.

P. Tóth Teodóra 2025. Oksági viszonyt jelölő kötőszavak korpuszalapú vizsgálata 14–15 éves diákok érvelő fogalmazásaiban. *Alkalmazott Nyelvtudomány, Különszám* 1: 237–255. https://dx.doi.org/10.18460/ANY.K.2025.1.013

Szilassy Eszter 2012. Az írás és fogalmazásjavítás stratégiái. *Anyanyelv-pedagógia* (5)1: 40–52. https://doi.org/10.21030/anyp.2012.1.5

Szentgyörgyi Rudolf 2017. A kétszintű magyar nyelv és irodalom érettségi vizsga 10 éve (2005–2015). *Anyanyelv-pedagógia* (10)4: 72–85. https://doi.org/10.21030/anyp.2017.4.5

Tóth Beatrix 2008. Fogalmazástanítás – miért és hogyan másképpen? *Anyanyelv-pedagógia* (1)1: 15–25. https://doi.org/10.21030/anyp.2008.1.2

Yaqub, Humaira – Tabassum, Rabia – Farooq, Mahwish 2016. Intra-rater reliability of holistic and rubric-based assessment of essay writing in Pakistan. *Scientific International (Lahore*) 28(4): 669–680.

(1) Magyar nyelv és irodalom részletes vizsgakövetelmények 2021. https://www.oktatas.hu/pub_bin/dload/kozoktatas/erettsegi/vizsgakovetelmenyek2024/magy_nyelv_es_irod_2024_e.pdf (2025. november 1.)

(2) Magyar nyelv és irodalom érettségi feladatlap 2024. https://dload-oktatas.educatio.hu/erettsegi/feladatok_2024tavasz_kozep/k_magyir_24maj_fl.pdf (2025. november 1.)

(3) Magyar nyelv és irodalom érettségi javítási-értékelési útmutató 2024. https://dload-oktatas.educatio.hu/erettsegi/feladatok_2024tavasz_kozep/k_magyir_24maj_ut.pdf (2025. november 1.)

Milán Juhász

An empirical examination of the reliability of writing assessment through the comparison of two assessment models

This study examines the reliability of writing assessment in the context of the Hungarian language and literature intermediate-level school-leaving examination, drawing on empirical data. The aim of the research was to explore the extent to which the marking and assessment guidelines introduced in 2024 provide an objective, consistent, and uniformly interpretable scoring framework for evaluating students' written texts, and to examine whether applying an alternative analytic assessment rubric can enhance inter-rater agreement. The empirical investigation involved forty Hungarian language and literature teachers with experience in administering the school-leaving examination, who assessed the same student's literary analytical essay using two different assessment systems. Interrater agreement and the internal consistency of the measurement instruments were analysed using multiple reliability indices, alongside an examination of the discriminative power of individual assessment criteria and differences in raters' scoring patterns. The findings indicate that within the assessment framework currently applied in the examination, inter-rater agreement is relatively low, and the interpretation of assessment criteria shows considerable variability. By contrast, the alternative analytic assessment model yields higher levels of agreement and a more stable assessment structure. The detailed performance

descriptors of the analytic rubric contribute to greater consistency in scoring decisions; however, the results also suggest that further refinement of the assessment dimensions would be justified. Overall, the study underscores the need for clearer specification of assessment criteria and the development of a more standardised approach to writing assessment.

Kulcsszók: anyanyelv-pedagógia, íráskészség, szövegalkotás, fogalmazásértékelés, értékelési megbízhatóság

Keywords: first-language pedagogy, writing skills, text production, writing assessment, assessment reliability

Az írás szerzőjéről

Juhász Milán
Eötvös Loránd Tudományegyetem, Budapest, Magyarország
juhasz.milan[kukac]btk.elte.hu
https://orcid.org/0009-0005-4526-2025