MAGYAR PEDAGÓGIA
113. évf. 3. szám 153–179. (2013)

A FOGALMAZÁSÉRTÉKELÉS MEGBÍZHATÓSÁGA KÉT FÜGGETLEN BÍRÁLÓ ÉRTÉKÍTÉLETEINEK ELEMZÉSE ALAPJÁN

Nagy Zsuzsanna
Szegedi Tudományegyetem, Oktatáselméleti Kutatócsoport

Az anyanyelvi kommunikációs képességek, köztük a megfelelően fejlett írásbeli szövegalkotás olyan eszköztudás, amit az élet minden területén alkalmazunk, és amelynek minősége alapvetően meghatározza tevékenységeink sikerességét. A külföldi neveléstudományi kutatásoknak (l. pl. *MacArthur*, *Graham* és *Fitzgerald*, 2008; *Hidi* és *Boscolo*, 2007; *Torrance*, *van Waes* és *Galbraith*, 2007; *Mäki*, 2002) erős és egyre hangsúlyosabbá váló, dinamikusan fejlődő iránya a fogalmazási folyamatoknak, illetve körülményeiknek, a fogalmazásképességnek, valamint az azt meghatározó tényezőknek a vizsgálata. Ez a megállapítás érvényes mind az anyanyelvi, mind az idegen nyelvi szövegalkotás kutatása terén. A hazai fogalmazáskutatás helyzetéről megállapítható, hogy bár egyre több hazai mérés (l. pl. *Molnár E. K.*, 2000, 2002, 2003, 2009; *Szilassy*, 2012; *Pintér*, 2009; *Nagy*, 2009) foglalkozik e képesség vizsgálatával, még ma is elhanyagolt területe az írásbeli szövegalkotás kutatása a magyar neveléstudományon belül. A fogalmazásképesség fejlettségének neveléstudományi mérések során megvalósuló, megbízható értékelése hozzájárul ahhoz, hogy objektív módszerekkel feltárjuk és megismerjük tanulóink szövegalkotási képességének szintjét, és megértsük, milyen tényezők befolyásolják a diákok szövegeinek minőségét.

Míg a tudásszintmérő és sok egyéb képesség fejlettségét vizsgáló tesztek esetében az értékelés megbízhatósága statisztikai módszerekkel biztosítható, az objektivitás megfelelő javítókulcsok segítségével elérhető, sőt számítógépes értékelés esetén az értékelők szubjektivitásából fakadó hibafaktor ki is küszöbölhető, addig a tanulói szövegek értékelése a körültekintően kidolgozott értékelési skálák ellenére is a feladatjavítók eltérő szigorúságából, illetve a skálapontok definícióinak eltérő értelmezéséből eredő értékelési hibáktól terhelt. Ennek következtében a fogalmazáskutatásokban az eredmények érvényességének biztosításában az alkalmazott értékelési szempontrendszer megbízható működése, a szövegeket bíráló szakértők megfelelő képzése központi szerepet játszik.

A tanulmányban a fogalmazásképesség értékelésének kérdéseivel foglalkozunk. A szakirodalom elemzése alapján bemutatjuk a papír alapú szövegértékelés korábbi hazai és külföldi kutatásokban alkalmazott módszereit és eszközeit, rávilágítunk az értékelés megbízhatósága kapcsán felmerülő problémákra. Ezt követően egy hazai és nemzetközi mintákon alapuló, saját fejlesztésű fogalmazás-értékelési szempontrendszer működésének vizsgálatára irányuló kutatás eredményeit mutatjuk be. Célunk az, hogy az értékelési

skálák működését (1) két független bíráló értékítéleteinek vizsgálatával, (2) a klasszikus és a valószínűségi tesztelmélet által kínált módszerekkel elemezzük, valamint (3) a bírálók értékelései közötti hasonlóságok és különbségek megismerése révén a szempontrendszer működésére vonatkozóan átfogó kép birtokába jussunk.

A kutatás elméleti háttere

A fogalmazásképesség meghatározásai

Az írott szöveg megalkotásának képességét az IEA 1980-as években zajlott nemzetközi fogalmazásvizsgálata (*Kádárné*, 1990; *Gorman*, *Purves* és *Degenhart*, 1988) mint „a nyelvi, gondolkodási és együttműködési képességnek, valamint olvasás- és írástechnikai készségeknek" (*Kádárné*, 1990. 19. o.) kommunikációs célok megvalósítása érdekében működő együtteseként írta le. *Kádárné* idézett meghatározása alapján a fogalmazásképesség egyrészt számos készség együttes működése által valósítja meg aktivitását, másrészt több más képességgel való egyidejű, egymást kiegészítő működése révén fejti ki hatását. Ennek megfelelően mind a hazai, mind a külföldi szakirodalomban fellelhető, hogy a fogalmazásképesség szerkezetét, működését modellező elméleti megközelítések összetett rendszerként tekintenek az írásbeli szövegalkotásért felelős képességre.

Nagy József (2002) személyiségmodelljében annak fejlődése szempontjából kiemelt szerepet kapnak a kommunikációs képességek, így a fogalmazás is. *Nagy* (1996) – összhangban a nemzetközi vizsgálatok értelmezésével – a fogalmazásképességet a kognitív kommunikációs képesség, ezen belül az írásképesség egyik – különféle képességekből, készségekből és egyéb komponensekből szerveződő – egységeként, „összefüggő gondolatrendszerek írásbeli közlésének szabályrendszereként" (64. o.) értelmezi. Ebbe a definícióba a szöveg megírásához szükséges anyaggyűjtéstől, rendszerezéstől, az adott műfajhoz és közlési helyzethez illeszkedő megformáláson át, az elkészült szöveg véglegessé alakításáig minden kognitív művelet beletartozik (*Molnár E. K.*, 2003).

A kognitív pszichológia klasszikus fogalmazásmodelljei a szövegalkotási képesség működését, fejlődését írják le (*Molnár E. K.*, 1996, 2003; *Eysenck* és *Keane*, 1997; *Beaugrande*, 1984; *Flower* és *Hayes*, 1980; *Hayes* és *Flower*, 1980; *Bereiter*, 1980; *Bereiter* és *Scardamalia*, 1987a, 1987b). *Molnár E. K.* (1996, 2003) kiemeli, hogy a klasszikus gondolkodás-lélektani modellek – az interaktív (*Beaugrande*, 1984), a rekurzív (*Flower* és *Hayes*, 1980; *Hayes* és *Flower*, 1980) és a képességintegrációs (*Bereiter*, 1980; *Bereiter* és *Scardamalia*, 1987a, 1987b) – közös vonása, hogy a jól írók (*expert*, *mature*) fogalmazási folyamatainak azonosításával a gyakorlatlan írók (*novice*, *immature*) problémáira világítanak rá. A kész szövegre fókuszáló fogalmazásszemlélettel szemben a kognitív pszichológiai irányzatok a szövegalkotási folyamatot állítják modelljük középpontjába, és azt összetett, többösszetevős rendszerként írják le, ahol a szöveg megalkotásának folyamata nem elszigetelten, hanem a feladatkörnyezettel (*Flower* és *Hayes*, 1980 idézi *Molnár E. K.*, 1996; *Hayes*, 1996) való folyamatos interakció mel-

lett zajlik, feltételezve például különböző szociális készségek és képességek működését is a fogalmazás közben.

A fogalmazások értékelésének szempontjai

Azt, hogy az írásművek értékelésekor milyen szempontokat érvényesítünk, meghatározza az, hogy a fogalmazásképességnek és a fogalmazási folyamatnak milyen elméleti modelljét, meghatározását fogadjuk el és tekintjük érvényesnek. A fogalmazásképességet, illetve a szöveg létrehozásának folyamatát modellező elméleti keretek elemzése révén a fogalmazási képesség szerkezetében azonosított összetevők, részképességek a fogalmazásértékelés során külön szempontokban minősíthetők.

Az iskolai osztályozás során megvalósuló fogalmazásértékelés jellemzően a fogalmazásképességnek csak néhány aspektusát minősíti. A pedagógusok általában egy vagy több – alapvetően két – szempontból, tartalmi, valamint nyelvtani-stilisztikai-helyesírási tekintetben, a hagyományos ötfokú skála mentén minősítik a tanulók szövegeit (*Molnár E. K.*, 2000), illetve egyre nagyobb arányban eltérő hosszúságú és részletezettségű szöveges értékelést adnak a diákok munkáiról (*Nagy*, 2011). Az ilyen típusú értékelés a fogalmazás két nagyobb, egymástól eltérő részképesség-csoportjának fejlettségét jellemzi: egyrészt a fogalmi kidolgozást, másrészt a különböző nyelvi konvenciók ismeretét (*Molnár E. K.*, 2000).

A fogalmazásképesség értékelésére irányuló neveléstudományi kutatások szintén ezt a kétirányú felosztást követik, ugyanakkor jellemzően nem egyszerűen e két szempont szerint minősítik a fogalmazásokat, hanem több részképességet értékelnek eltérő számú szempont mentén (*Molnár E. K.*, 2000). A fogalmazáskutatások szempontrendszereinek kidolgozása komoly körültekintést igényel. El kell döntenünk, tárgyunk mely aspektusát kívánjuk feltárni, mire fektetjük a hangsúlyt az értékelésnél, és ennek megfelelően kell összeállítanunk a feladatokat. Szempontunk lehet, hogy a tanulók által alkotott szövegek mennyire hatékonyak, de vizsgálhatjuk azt is, hogy a felmérés résztvevői mennyire tudják a művelt köznyelvnek, a kért műfaj sajátosságainak megfelelő módon megfogalmazni gondolataikat. A hazánkban zajlott országos reprezentatív fogalmazáskutatások elsősorban ez utóbbira voltak kíváncsiak és ezt egészítették ki egyéb szempontokkal (*Molnár E. K.*, 2003).

A szövegalkotási képesség mérésére irányuló vizsgálatok eszközei lehetnek a különböző kontextusba helyezett fogalmazásfeladatok, melyek megoldásakor a tanulóknak meghatározott szövegtípusoknak megfelelő szöveget vagy szövegeket kell alkotniuk meghatározott idő alatt. Ezeket – a vizsgálatok függő változójának, a szövegalkotási képesség mérésére készült – eszközöket egészíthetik ki a háttéradatok, illetve más, az eredményeket feltehetően befolyásoló tényezők feltérképezésére készített kérdőívek, tesztek (*Kádárné*, 1990). A fogalmazási képesség méréséhez használt feladatok kiválasztásakor figyelembe kell vennünk, hogy a választott módszerek jelentősen befolyásolják azt, hogy ennek az összetett képességnek mely komponensei kerülnek a felszínre, mi válik valójában mérhetővé a feladatok értékelése által (*Molnár E. K.*, 2000). Emellett *Purves* (1992) alapján *Molnár Edit Katalin* (2000) felhívja a figyelmet arra is, hogy a feladatmegoldás körülményei szintén hatást gyakorolnak arra, hogy a szövegalkotási ké-

pesség ki tud-e bontakozni a maga teljes valójában, a tanuló számot tud-e adni írásbeli kommunikációjának fejlettségéről, hiszen a fogalmazás folyamata, üteme egyéni, amely nem feltétlenül működik a diákra egyébként jellemző módon a mérési alkalom egy-két órája során.

A kutatások során összegyűjtött tanulói szövegek értékelésére a szervezők különböző szempontrendszereket dolgoznak ki, melyek segítségével egységesen pontozhatják a született dolgozatokat. A következőkben néhány hazai és nemzetközi példán mutatjuk be a fogalmazásértékelési szempontrendszerek típusait.

Az IEA 1980-as években lebonyolított vizsgálatánál (Written Composition Study 1984–1985) például *Takala* (1988, magyarul ismerteti *Kádárné*, 1990; *Molnár E. K.*, 2003) írásbeli kifejezőképesség szerveződését bemutató elméleti modelljére alapozott szempontrendszerrel értékelték a tanulói fogalmazásokat. *Takala* háromszintű modelljében a szövegtervezésben, illetve a szövegszerkesztésben való jártasság két-két készségcsoportra – az intellektuális készségek és a társas-együttműködési készségek, valamint a nyelvi készségek és az írás- és szövegelrendezés készség körére – tagolódik. Ezek az ismeretek hat csoportját foglalják magukban: (1) fogalmak, gondolkodásmódok ismerete és szókincs; (2) anyagszervezési és -szerkesztési elvek; (3) kommunikációs normák, kifejezésmódok, stíluseszközök ismerete; (4) nyelvtani, nyelvhelyességi, írásjel-használati ismertek (5) helyesírási szabályok; (6) írásjegyek, jelrendszerek, írásmódok, formaszabályok ismerete (*Kádárné*, 1990). A fogalmazásvizsgálat eszközrendszerének kidolgozásakor az ismeretek e hat köréhez rendeltek értékelési kritériumokat. Ennek alapján az IEA-vizsgálat szempontrendszere hat analitikus és egy holisztikus szempontot tartalmaz, vagyis (1) a tartalom, (2) a felépítés/szerkezet, (3) a stílus, (4) a nyelvhelyesség, (5) a helyesírás és (6) a külső alak mellett a bírálók egy összbenyomás-osztályzattal is minősítették a tanulók írásait. Munkájukat a nemzetközi etalonskálák, mintadolgozatok is segítették. Az elkészült tanulói szövegeket minden esetben két bíráló értékelte. A végleges osztályzatokat zsűrimódszerrel alakították ki, kettejük folyamatos egyeztetése révén konszenzusosztályzatokat állapítottak meg (*Kádárné*, 1990).

Orosz (1972) 18 szempont alapján értékelte a vizsgálata eredményeként készült szövegeket. A tartalmat négy további nézőpontból elemezte, így figyelt az anyaggyűjtésre, a fogalomválasztásra, valamint az ítéletalkotásra és -kapcsolásra is. A szerkesztést minősítő hét osztályzatban a szöveg globális megalkotástól egészen a szavak szintjéig vizsgálta és értékelte az írásokat, a stilizálás minősítésének pedig szintén hét szempontot szentelt.

Horváth (1998) 32 szempontot alkalmazott mérésében; a tartalom, szerkezet és nyelvi megformálás mellett az összbenyomás, a helyesírás, a külalak, a hatékonyság és a feladatspecifikáció szerint is értékelte a fogalmazásokat. A tartalmat minősítő tíz szempont között külön szerepel például a gondolatok és az álláspont kifejtettségének, az érvek alkalmazásának, a hitelességnek és a hatékonyságnak az értékelése. A szerkezeti jegyek között a gondolategységek elkülönítése mellett az egész szöveg, illetve a kisebb szövegegységek felépítését értékelő szempontok is megjelennek. A nyelvi megformáláson belül a hangnem, a stílus, a szókincs és a nyelvhelyesség megfelelését minősítő szempontok találhatóak.

Molnár Edit Katalin (2000, 2002, 2003) vizsgálatának fogalmazásértékelési szempontrendszere részben az IEA-mérés (*Kádárné*, 1990) skáláiból, részben egy más terüle-

teket – alapvetően írás- és helyesírási készséget – mérő vizsgálat (*Vidákovich*, 1990) szempontjaiból merített. A fogalmazások értékelése egy holisztikus és több analitikus szempont mentén történt. Míg a holisztikus szempont a tartalmi és a formai jegyek szétválasztása nélküli minősítette a tanulók munkáit, addig az analitikus szempontok egyes szövegjellemzőkre fókuszáltak. A tartalom, a szerkezet/műfaj/felépítés, illetve a stílus értékelésekor az IEA kritériumait vette figyelembe, a nyelvhelyességen belül a tanuló nyelvhasználatát, annak szabadosságát, világosságát, egyértelműségét jellemezte, a diákok munkáinak külalakját pedig a jelek elrendezettsége, a szöveg megjelenítésének szépsége szerint minősítette *Vidákovich* (1990) kategóriái alapján.

A bemutatott szövegvizsgálatok egymáshoz hasonló szempontrendszerrel dolgoznak. A különbségek abból adódnak, hogy a szempontokat különböző alszempontokra bontják, és így a három központi értékelési egység – a tartalom, a szerkezet és a stílus – egyes műveleteit külön osztályozzák (*Molnár E. K.*, 2003).

A nemzetközi szakirodalomban alkalmazott fogalmazás-értékelési szempontrendszerek három csoportját különíthetjük el, melyeket az alábbiakban néhány példa bemutatásával szemléltetünk. A vizsgálatok egy része egy globális osztályzattal minősíti a mérés során nyert tanulói fogalmazásokat. Ezek az egyszerű értékelési rendszerek, hasonlóan a hagyományos iskolai értékeléshez, a szövegeket egy értéklettel helyezik el a teljes mintában, ezzel fejezve ki a közöttük megfigyelhető minőségi különbségeket. *Davis* (2005) a nyelvtani tesztelés szövegalkotási teljesítményre gyakorolt hatását elemezte, főiskolás hallgatók szövegeit négyfokú holisztikus értékeléssel minősítette. *Chai* (2006) elsősorban a vizsgált tanulók szövegalkotás előtt készített vázlatainak minőségére, annak hatásaira figyelt mérésében, összefüggést keresett a vázlatok és az elkészült végleges szövegek minősége között. Ebben a mérésben a fogalmazások értékelése egy holisztikus osztályzattal történt. *Gelati* és *Boscolo* (2009) négyfokú holisztikus értékelést alkalmazott általános iskolás tanulók körében végzett kísérletében.

Több vizsgálat dolgozik az IEA-mérés ismertetett szempontrendszeréhez hasonló értékelési skálákkal. Ezekre a fogalmazáskutatásokra az jellemző, hogy szempontrendszerükben az IEA-szempontok között is megtalálható kritériumok jelennek meg, azokkal részben vagy egészében megegyező szempontokat tartalmaznak, esetleg finomítva, alszempontokra bontva egyes elemeket. *Engelhardt*, *Gordon* és *Gabrielson* (1991) 18 írásfeladatot értékelt tartalom és szerkezet, stílus, mondatforma, szóhasználat és nyelvtan szerint. *Zhang* és *Vukelich* (1998) az írás előtti tevékenységek (*prewriting*) hatásait elemző kutatásában egy holisztikus és öt analitikus szempontból értékelte a fogalmazásokat. Háromfokú skálán minősítette a mondatszerkezetet, a nyelvtant, a szóhasználatot, a kifejtést és az elrendezést. *Popp*, *Ryan*, *Thompson* és *Behrens* (2005) általános iskolás tanulók körében végzett mérésükben az ötletek, az elrendezés, a hangnem, a szóhasználat, a mondatok gördülékenysége és a nyelvtani szabályok betartása alapján jellemezték a diákok fogalmazásainak minőségét. *Barkaui* (2007) szintén az IEA szempontjaihoz hasonló értékeléssel dolgozott, egy holisztikus és öt analitikus – tartalom, szerkezet, nyelvtan, írástechnika, stílus – szempont szerint jellemezte a szövegeket. *Crawford* és *Smolkowski* (2008) a fogalmazásokat három aspektusból – stílus és gördülékenység, tartalom és szerkezet, illetve nyelvhasználat szerint – négyfokú skálán minősítette.

A vizsgálatok egy másik hányada olyan szempontrendszert alkalmaz, amely vagy speciális szövegtípusok értékelését, vagy a szövegek árnyaltabb jellemzését teszik lehetővé újabb szempontok bevonásával. *Gearhart*, *Herman*, *Novak* és *Wolf* (1995) a nagymintás fogalmazásmérések értékelésrendszerének oktatási hatékonyságát vizsgálva az IEA-mérés rendszeréből is ismerős szempontok (fókusz/elrendezés, kidolgozás, írástechnika), valamint általános kompetencia és a *Writing What You Read* (WWYR) többszintű keretrendszer szempontjai alapján értékelték a tanulók szövegeit. A WWYR-rendszer egy holisztikus és öt analitikus dimenziót tartalmaz a tanulók narratív szövegeiben megjelenő téma, karakter, elrendezés, cselekmény és kommunikációbeli fejlődésének hatfokú értékelésére. A szövegalkotás médiumának, többek között, a szövegek minőségére gyakorolt hatásait vizsgálva *Whithaus*, *Harrison* és *Midyette* (2008) főiskolás hallgatók fogalmazásait négyfokú skálán értékeltette a vizsgálatban részt vevő bírálókkal a téma, az érvelés, a mondatszerkezet és a szóhasználat, logikai konzisztencia és a nyelvtani hibáktól való mentesség szerint.

Ransdell, *Levy* és *Kellogg* (2002), illetve *Ransdell* és *Levy* (1996) a fogalmazások értékelésére két független bírálót és komplex értékelési rendszert alkalmazott. A hat alcsoportos minőségi skála (*Six-Subgroup Quality Scale*) a szóhasználat és elrendezés, a technikai minőség, a tartalom, a szándék/közönség/hangnem, a szerkesztés és kidolgozás, valamint a stílus szempontjából minősíti a szövegeket, és ezeket az alcsoportokat bontja tovább, összesen 13 szemponttá a fogalmazások különböző aspektusainak minősítése érdekében. *Segev-Miller* (2004) tanárjelöltek fogalmazással kapcsolatos metakognitív stratégiáinak vizsgálatakor a résztvevők fogalmazásainak értékeléséhez több szempontot alkalmazott. Értékelte a témát, a makropropozíciót, a kidolgozást, a megfelelő retorikai struktúrát, az explicit kohéziót, a nyelvi megformálást, az értelmezést és az idézést.

Ezek az utóbbi vizsgálatok szempontrendszereikkel részletesebb fogalmazásértékelést valósítanak meg, illetve egy-egy kiválasztott szövegtípus vizsgálatát teszik lehetővé. A hagyományos iskolai fogalmazásértékelés során többnyire a tanulói szövegek holisztikus értékelése érvényesül és mellőzött az egyes szövegjellemzők minősítése. Ezen egyszerű és a fogalmazáskutatások egy része által is használt skáláknak az az előnye, hogy a szövegek gyors értékelését teszik lehetővé, és megvalósítják azt a kritériumot, hogy az egyes munkákat minőségük alapján elhelyezzük a teljes szövegkorpuszban. Ugyanakkor a globális értékelés nem biztosítja a fogalmazások különböző aspektusokból történő értékelését. A holisztikus értékelés nem képes arra sem, hogy a fogalmazásképesség fejlettségét az egyes szövegjellemzők mentén külön minősítse, ezáltal rávilágítson arra, hogy a tanuló a szövegalkotás közben egyes vizsgált tényezőket jobban, míg másokat gyengébben kezel. Kívánatos tehát a fogalmazások többszempontú minősítése az árnyalt szövegértékelés érdekében.

A fogalmazás-értékelési szempontrendszerek megbízhatósága és működésük statisztikai jellemzése

A fogalmazásvizsgálatokban különös problémát jelent a megbízhatóság biztosítása. A szövegértékelés eredményeinek érvényessége nagymértékben függ az alkalmazott értékelési skálapontok egyértelmű definiálásától, ugyanakkor ez nem biztosítja maradékta-

lanul a szempontrendszer megbízható működését. A skálapontok körültekintő megfogalmazása ellenére jellemző, hogy a bírálók eltérően értelmezik az értékkategóriákat. Ebből következik, hogy időnként ugyanazt a szöveget a különböző értékelők eltérően minősítik. *Vígh* (2010) hasonló problémákra hívja el a figyelmet az idegen nyelvi íráskészség pontozásának kérdéseit elemezve. Több – *Vígh* (2010) által is hivatkozott – nemzetközi kutatás (*Weigle*, 1998; *Engelhard* és *Myford*, 2003; *Eckes*, 2005, 2008; *Schoonen*, 2005) eredménye alapján a vizsgázók szövegeinek értékelésekor a bírálók több területen – például szigorúságukban, az értékelési szempontrendszer alkalmazásában, valamint a gyengébb és jobb teljesítményű diákok teljesítménye pontozásának következetességében – eltéréseket mutatnak. Az említett jelenségek hátterében több ok is állhat. *Szilassy* (2012) *Hillocksra* (1986) hivatkozva kiemeli, egy szöveg minősége nehezen fordítható le számokkal kifejezett értékletekké. A fogalmazások bírálóinak gondolkodása kulturálisan meghatározott. Emellett saját személyiségjegyeik, tulajdonságaik (*Molnár E. K.*, 2000; *Szilassy*, 2012), valamint bizonyos mértékben az értékelendő szöveg tartalmára, az általa megvalósított kommunikációs helyzetre vagy általában a szövegalkotásra, a szövegek minőségére vonatkozó tudásuk (*Nagy*, 2009), sőt a szerzőről való korábbi ismereteik (*Horváth*, 1998; *Szilassy*, 2012) is befolyással bírnak mások szövegeiről alkotott értékítéleteikre. Ennek következtében a fogalmazásértékelés mindig valamilyen mértékű bírálói szubjektivitástól terhelt.

A fogalmazás-értékelési szempontrendszerek működését jellemző egyik mutató a reliabilitást kifejező *Cronbach*-α érték, ami a fogalmazásvizsgálatokban 0,75 felett már elfogadható. A holisztikus szempontra mint függő változóra végzett regresszióanalízis segítségével megadható az analitikus szempontok által megmagyarázott összes variancia, vagyis megállapítható, hogy az értékelők szövegről alkotott globális ítéletét mennyiben határozzák meg az analitikus szempontokban kifejezett szövegjellemzők, illetve értékítélete mennyiben tulajdonítható más, külső tényezőknek.

A fogalmazáskutatásokra jellemző, hogy a különböző szempontokhoz tartozó skálapontok pontos definiálása mellett az egyes tanulói dolgozatok több bíráló által történő értékelése révén igyekeznek növelni a mérés megbízhatóságát. Ilyen módon lehetőség nyílik arra, hogy az értékeléshez használt szempontrendszereket a Cronbach-α értékek és a regresszióelemzés mellett az értékelők ítéletei közötti korrelációkkal, illetve a *Kendall*-féle konkordanciaelemzés eredményeivel (pl. *Beyreli* és *Ari*, 2009) jellemezzék. A konkordanciaértékek az utóbbi elemzésekben 0 és 1 közé eső számok, melyek az értékelők közötti egyetértés mértékét fejezik ki. Minél nagyobb ez az érték, annál inkább azonos ítéleteket alkotnak a bírálók az egyes szempontok szerint.

Az értékelők ítéletei közötti korrelációkat (r) a különböző bírálók azonos szempontokra adott értékleteinek összefüggései jelentik, és minél erősebbek ezek a korrelációs együtthatók, annál nagyobb az összhang az értékelők ítéletei között. *Isonio* (1991) vizsgálatában például ez az érték 0,76, *Beyreli* és *Ari* (2009) mérésben 0,66–0,83 között volt. Az IEA 1980-as évek végén végzett nemzetközi fogalmazáskutatásában is elemezték az értékelők ítéletei közötti korrelációk erősségét, melyek az összbenyomás osztályzatok esetén 0,61 és 0,82 közötti értékeket mutattak, az összefüggések közepesek, illetve erősek voltak (*Gorman*, *Purves* és *Degenhart*, 1988; *Kádárné*, 1990).

A valószínűségi tesztelmélet széles körű elemzési lehetőségeket kínál a képességkutatások, így a fogalmazásvizsgálatok számára is (*Molnár Gy*., 2003, 2005, 2006, 2008, 2013; *Vígh*, 2010). Ennek ellenére csak kevés olyan hazai kutatást ismerünk, amely ilyen típusú vizsgálatokra vállalkozott volna a nyelvi képességek terén (pl. *Vígh*, 2008, 2010; *Kontra*, 2009; *Molnár* és *Józsa*, 2006; *Dávid*, 2008). Az írásbeli szövegalkotás vizsgálatait tekintve is megállapítható, hogy a parciáliskredit-modell nyújtotta lehetőségek segítségével a korábbinál szélesebb eszköztár áll rendelkezésünkre mind a fogalmazásképesség, mind az ennek értékelését lehetővé tevő szempontrendszerek működésének jellemzésére. Ugyanakkor nem tudunk olyan hazai kutatásról, amely a parciáliskredit-modellel történő elemzések előnyeit kihasználva vizsgálta volna a tanulók szövegalkotási képességét, vagy tesztelte volna az alkalmazott mérőeszköz megbízhatóságát. Ugyanakkor a nemzetközi anyanyelven vagy idegen nyelven írt szövegek íratása által vizsgálódó fogalmazáskutatásokban (pl. *Engelhard*, 1994; *Gyagenda* és *Engelhard*, 1998; *Griffin* és *Anh*, 2005; *Sugita*, 2009; *Barkaoui*, 2011; *Sudweeks*, *Reeve* és *Bradshaw*, 2004; *Wiseman*, 2012) az értékelők szigorúságát, a szempontrendszerek működését az eredmények *Rasch*-modellel történő elemzése révén is vizsgálják.

A nem dichotóm adatok, így a többfokú skálákon történő fogalmazásértékelések eredményeinek elemzésére alkalmas parciáliskredit-modell lehetővé teszi, hogy a szövegek minőségének jellemzéséhez használt értékelési szempontokat, skáláik működését jellemezzük (*Molnár Gy.*, 2008, 2013). Mivel a parciáliskredit-modell közös skálán helyezi el a tanulók képességszintjét és az egyes szempontok nehézségi paramétereit, a vizsgált tanulók képességszintje alapján határozhatjuk meg a szempontok átlagos nehézségét. Ugyancsak vizsgálható az értékelési szempontok modellilleszkedése, a skálapontok egymástól való elkülönülése, a mintának való megfelelése, illetve az értékelők skálahasználata is. A *Rasch*-modellel történő elemzések több értékelő munkájának jellemzését, a többaspektusú modell segítségével szigorúságuk összehasonlítását szintén lehetővé teszik (*Vígh*, 2010).

A vizsgálat módszerei és eszközei

Kutatásunk során 8. évfolyamos tanulóktól (N=429) az elbeszélés műfajában kértünk szövegalkotást. Az adatfelvételre a 2010–2011-es tanév első félévében került sor. Minden tanulói szöveget két független bíráló értékelésének vetettünk alá. A fogalmazások értékelői a szövegek minősítését megelőzően pedagógiai mérés-értékelés területén szereztek képzettséget, rendelkeztek értékelői, feladatjavítói tapasztalattal, illetve tájékoztatást kaptak az alkalmazott fogalmazás-értékelési szempontrendszerről, próbaértékelés során tanulmányozták annak skáláit.

A szövegek minőségének megítéléséhez egy tíz szempontból álló, saját fejlesztésű értékelési rendszert használtunk, ami egy holisztikus és kilenc analitikus szempontból épült fel (*Gorman*, *Purves* és *Degenhart*, 1988; *Kádárné*, 1990; *Molnár E. K.*, 2003; *Ransdell* és *Levy*, 1996). Az értékelőktől azt kértük, hogy hazai és nemzetközi mintákon alapuló szempontrendszerünk mentén, ötfokú skálán – (1) tartalom, (2) szövegtípus, il-

letve (3) hangnem szerinti feladattartás, (4) szerkezet és kidolgozás, (5) stílus, (6) érthetőség, (7) nyelvhelyesség, (8) helyesírás, központozás, (9) külalak és olvashatóság – szerint minősítsék a tanulók munkáit. Mindemellett szintén ötfokú, globális értéklettel fejezzék ki az egyes dolgozatokról alkotott általános ítéletüket.

A tartalom értékletben az értékelők az anyaggyűjtést, a mondanivalót, a hitelességet, relevanciát, a tartalom mélységét és gazdagságát minősítették. A feladattartás szempontokban került sor arra, hogy a tanulóktól kért szövegtípust és a feladatnak megfelelő hangnemet vizsgáljuk. A szerkezet és kidolgozás skálája nem csak azt vizsgálta, hogy milyen a szöveg makroszintű elrendezése, a szöveg felépítésének logikája, hanem értékelte a szöveg belső koherenciáját, a bekezdések felépítését és egymáshoz való kapcsolódását is. A stílussal a nyelvi kifejezést, megformálást, a választékosságot minősítettük, de elválasztottuk ettől a feladathoz illeszkedő hangnem kiválasztását és megtartását. Az érthetőség szempontban a szöveg megértésre való előkészítettségét, a megfogalmazás világosságát vizsgáltuk. A nyelvhelyesség szempontban az írott köznyelv grammatikai szabályainak betartását értékeltük. A helyesírást és a központozást egy szemponton belül értékelték. A külalakkal egy jegyben vizsgáltuk az olvashatóságot, itt néztük meg az íráskép rendezettségét, a jelek elrendezését, a szöveg megjelenítésének szépségét, a szótávolságok és betűnagyságok arányosságát, egyenletességét. Valamennyi szemponthoz ötfokú értékelési skálát dolgoztunk ki, melyben röviden leírtuk azokat a szövegjellemzőket, amelyek a tanulók által létrehozott szövegekben megfigyelhetők.

A mérőeszköz értékelési skáláit korábbi, 2010 tavaszán 4. és 8. évfolyamos tanulók körében megvalósított fogalmazásvizsgálatunk (*Nagy*, 2010) alkalmával egy bíráló értékleteinek klasszikus tesztelméleti módszerekkel történő elemzése által már teszteltük. A kvantitatív elemzés eredményei szerint a szempontrendszer reliabilitása mindkét vizsgált évfolyamon megfelelő (Cronbach-α_1=0,94, Cronbach-α_2=0,96). Az összbenyomás értékletre mint függő változóra végzett regresszióanalízis eredménye szerint az analitikus szempontok körülbelül 91%-ban magyarázták a szövegek minőségéről átfogó jellemzést adó holisztikus osztályzatot, és valamennyi szemponton nyújtott teljesítmény szignifikáns ($p<0,01$) összefüggést mutatott az összbenyomás-jeggyel és a többi szempont szerint tapasztalt teljesítménnyel is. Eredményeink alapján a szempontrendszer működését kielégítőnek, az értékelési skálákat további használatra alkalmasnak találtuk (*Nagy*, 2010).

Jelen kutatásunkban a szempontrendszer működését további vizsgálatnak vetettük alá. Mivel ugyanazon tanulói minta fogalmazásainak két független bíráló által történő értékeléseivel rendelkeztünk, lehetőségünk nyílt arra, hogy az értékelők ítéleteinek összevetésével további következtetéseket fogalmazzunk meg az alkalmazott fogalmazásértékelési rendszer megbízhatóságára vonatkozóan. Vizsgálatunkban a klasszikus tesztelmélet módszerei mellett alkalmaztuk a parciáliskredit-modell nyújtotta eszköztárat is. Az adatok elemzését az SPSS-programcsomag és a ConQuest (*Wu*, *Adams* és *Wilson*, 1998) elemzőszoftver segítségével végeztük el.

A vizsgálat eredményei

A szempontrendszer működésének vizsgálatára irányuló kutatás eredményeinek bemutatásakor előbb a klasszikus, majd a valószínűségi tesztelmélet módszereivel végzett statisztikai elemzéseket ismertetjük. A klasszikus tesztelmélet eszközrendszerére alapozó analízisek között kitérünk a két értékelő ítéletei alapján külön számított reliabilitásértékekre, majd bemutatjuk az egyes szempontokra adott értékletek belső összefüggésrendszerét. Ezt követően a két értékelő azonos szempontokból alkotott ítéleteinek korrelációit közöljük, végül az összbenyomás osztályzatokra végzett regresszióanalízisek eredményeit elemezzük. Az ezt követő fejezetben a valószínűségi tesztelméleti módszerekkel végzett vizsgálatok, így a parciáliskredit-modellel, illetve a többaspektusú modellel folytatott elemzéseink eredményeit foglaljuk össze.

A szempontrendszer működése a klasszikus tesztelmélet modelljei alapján

A szempontrendszer működésének klasszikus tesztelméleti eszközökkel történő vizsgálatát a reliabilitásmutató meghatározásával kezdtük. Az összbenyomás szempont az értékelők szövegről alkotott általános benyomását fejezi ki, és ezt – ahogyan az a regresszióanalízisek eredményeinek bemutatásakor látható – az analitikus szempontokban kifejezett szövegjellemzők határozzák meg leginkább. Emiatt a globális skálát kihagytuk a reliabilitás vizsgálatából.

Az analitikus szempontok elemzése alapján a két bírálóra vonatkozóan két Cronbach-α értéket állapítottunk meg, melyek között csak néhány ezrednyi a különbség. A szempontrendszer megbízhatósági mutatója mindkét bíráló esetén megfelelőnek bizonyult, mindkét értékelőnél a Cronbach-α_1 0,95.

Az 1. és a 2. táblázat a bírálók értékítéleteink szempontonkénti korrelációit foglalja össze. Az eredmények alapján – mind az első, mind a második bíráló értékleteit vizsgálva – valamennyi értékelési szempont szignifikánsan ($p<0,001$) korrelált minden más értékelési szemponttal. A korrelációs együtthatók erősségének mintázata a két értékelő esetén hasonló. A tartalom, a feladattartás: szövegtípus, illetve hangnem, a szerkezet és kidolgozás, a stílus, az érthetőség és a nyelvhelyesség szempontok egymással való összefüggései bizonyultak erősebbnek, míg a helyesírás, központozás, valamint a külalak és olvashatóság osztályzatok gyengébben korreláltak a többi szemponttal. Az első bírálónál az utóbbi két jegy 0,37 és 0,65, a másodiknál 0,53 és 0,76 közötti erősségű összefüggéseket mutatott a szempontrendszer többi skálájával, míg az egyéb szempontok korrelációs együtthatói 0,73 és 0,95, illetve 0,69 és 0,84 közötti értékeket vettek fel. Ez a két szempont a korrelációk elemzése alapján a többitől némiképp függetlenebbnek bizonyult.

A helyesírás, központozás, illetve a külalak és olvashatóság skála az értékelők holisztikus ítéletével is a többi szempontnál gyengébb összefüggést mutatott. Az első bírálónál ezek az osztályzatok 0,60 és 0,40, a másodiknál 0,68 és 0,56 erősséggel korreláltak az összbenyomással. Ugyanakkor a többi szempont esetén mindkét értékelőnél magasabb, 0,78 és 0,91, illetve 0,81 és 0,88 közötti korrelációs értékeket tapasztaltunk.

1. táblázat. Az első bíráló értékítéleteinek szempontonkénti összefüggései

Szempontok	*T*	*Fsz*	*Fh*	*SzK*	*S*	*É*	*Ny*	*HK*	*KO*
Tartalom	–								
Feladattartás: szövegtípus	0,77	–							
Feladattartás: hangnem	0,85	0,88	–						
Szerkezet és kidolgozás	0,85	0,86	0,95	–					
Stílus	0,78	0,78	0,82	0,81	–				
Érthetőség	0,82	0,81	0,85	0,86	0,85	–			
Nyelvhelyesség	0,75	0,73	0,79	0,79	0,91	0,83	–		
Helyesírás, központozás	0,57	0,57	0,59	0,61	0,62	0,63	0,65	–	
Külalak és olvashatóság	0,37	0,40	0,39	0,39	0,45	0,43	0,44	0,42	–
Összbenyomás	0,91	0,80	0,87	0,88	0,81	0,84	0,78	0,60	0,40

Megjegyzés: minden összefüggés $p<0,001$ szinten szignifikáns; T=tartalom; Fsz=szövegtípus szerinti feladattartás; Fh=hangnem szerinti feladattartás; SzK=szerkezet és kidolgozás; S=stílus; É=érthetőség; Ny=nyelvhelyesség; HK=helyesírás és központozás; KO=külalak és olvashatóság.

2. táblázat. A második bíráló értékítéleteinek szempontonkénti összefüggései

Szempontok	*T*	*Fsz*	*Fh*	*SzK*	*S*	*É*	*Ny*	*HK*	*KO*
Tartalom	–								
Feladattartás: szövegtípus	0,78	–							
Feladattartás: hangnem	0,78	0,84	–						
Szerkezet és kidolgozás	0,80	0,84	0,82	–					
Stílus	0,77	0,76	0,79	0,78	–				
Érthetőség	0,80	0,75	0,77	0,80	0,83	–			
Nyelvhelyesség	0,72	0,71	0,76	0,69	0,81	0,74	–		
Helyesírás, központozás	0,62	0,60	0,66	0,60	0,64	0,64	0,76	–	
Külalak és olvashatóság	0,51	0,48	0,53	0,50	0,54	0,54	0,58	0,53	–
Összbenyomás	0,86	0,85	0,88	0,87	0,86	0,86	0,81	0,68	0,56

Megjegyzés: minden összefüggés $p<0,001$ szinten szignifikáns; T=tartalom; Fsz=szövegtípus szerinti feladattartás; Fh=hangnem szerinti feladattartás; SzK=szerkezet és kidolgozás; S=stílus; É=érthetőség; Ny=nyelvhelyesség; HK=helyesírás és központozás; KO=külalak és olvashatóság.

Az első és a második bíráló által azonos szempontokra adott értékletek összefüggéseit a 3. táblázatban közölt korrelációs együtthatók alapján követhetjük nyomon. Valamennyi analitikus szempont és az összbenyomás esetén is a két bíráló ítéletei erősen ($0,85<r<0,93$) és szignifikánsan korrelálnak egymással. A magas r-értékek arra utalnak,

hogy a bírálók közel azonosan ítélik meg a tanulói szövegeket, bírálataik szoros összefüggésben állnak egymással valamennyi vizsgált jellemző mentén.

3. táblázat. A két bíráló azonos szempontokra adott értékítéleteinek összefüggései

Szempontok	*Összefüggések (r)*
Tartalom	0,85
Feladattartás: szövegtípus	0,86
Feladattartás: hangnem	0,89
Szerkezet és kidolgozás	0,87
Stílus	0,88
Érthetőség	0,85
Nyelvhelyesség	0,88
Helyesírás, központozás	0,93
Külalak és olvashatóság	0,86
Összbenyomás	0,85

Megjegyzés: minden összefüggés szignifikáns $p<0,001$ szinten.

Azonban az összbenyomás-teljesítményre mint független változóra végzett regresszióelemzések eredményei (4. táblázat) már különbségeket mutatnak a két bíráló értékelési teljesítményében. Bár az analitikus szempontok 88, illetve 90%-ban magyarázzák a holisztikus értékletek varianciáját, az egyes szempontok hozzájárulása az összbenyomáshoz igen eltérő. Ha megvizsgáljuk az első értékelő értékítéleteinek jellemzőit, akkor szembetűnő, hogy az összbenyomás varianciájának csaknem felét a tartalom határozza meg, és a többi szempont közül csak a szerkezet és kidolgozás hozzájárulása jelentősebb, a többié elenyésző. Az értékek szignifikanciáját vizsgálva szintén azt tapasztaltuk, hogy statisztikailag releváns mértékben csak ez a két szempont járul hozzá az értékelő szövegről alkotott általános benyomásához.

A második értékelő esetén (5. táblázat) sokkal kiegyenlítettebbek az értékek. A bíráló ítéletét legnagyobb mértékben a feladattartás: hangnem határozza meg, viszont a többi szempont magyarázóereje is kielégítő, a helyesírás, központozás, illetve a külalak és olvashatóság kivételével szignifikáns is.

A klasszikus tesztelméleti módszerekkel végzett elemzések eredményeinek összegzéseként megállapítható, hogy bár mindként bíráló esetén az egyes szempontok erős, illetve közepesen erős összefüggését tapasztaltuk, illetve az azonos szempontokra adott ítéletek között erős korrelációkat találtunk, a két értékelő bírálói gondolkodása nem ekvivalens. Az összbenyomásra végzett regresszióelemzések eredményei a két értékelő munkája között jelentős különbségeket jeleztek, az analitikus szempontok globális ítélethez való hozzájárulásának igen nagy különbségeit mutatták ki.

4. táblázat. Az összbenyomás osztályzatra mint függő változóra végzett regresszióanalízis eredménye az első bíráló esetén

Szempontok	*r*	*β*	*r*β*100*
Tartalom	0,91	0,51	45,70*
Feladattartás: szövegtípus	0,80	0,05	3,84
Feladattartás: hangnem	0,87	-0,07	-5,85
Szerkezet és kidolgozás	0,88	0,32	28,39*
Stílus	0,81	0,08	6,37
Érthetőség	0,84	0,06	5,05
Nyelvhelyesség	0,78	0,04	2,81
Helyesírás, központozás	0,60	0,02	1,16
Külalak és olvashatóság	0,40	0,01	0,18
Összes megmagyarázott variancia			87,65

Megjegyzés: Minden r-érték p<0,001 szinten szignifikáns. *A β-érték p<0,001 szinten szignifikáns.

5. táblázat. Az összbenyomás osztályzatra mint függő változóra végzett regresszióanalízis eredménye a második bíráló esetén

Szempontok	*r*	*β*	*r*β*100*
Tartalom	0,86	0,18	15,13**
Feladattartás: szövegtípus	0,85	0,10	8,43*
Feladattartás: hangnem	0,88	0,20	17,54**
Szerkezet és kidolgozás	0,87	0,16	14,19**
Stílus	0,86	0,14	12,27**
Érthetőség	0,86	0,15	12,44**
Nyelvhelyesség	0,81	0,10	7,75*
Helyesírás, központozás	0,68	0,02	1,28
Külalak és olvashatóság	0,56	0,02	0,94
Összes megmagyarázott variancia			89,96

Megjegyzés: Minden r érték szignifikáns p<0,001 szinten.
*A β érték szignifikáns p<0,05 szinten. **A β szempont magyarázó ereje szignifikáns p<0,001 szinten.

A szempontrendszer működése a valószínűségi tesztelmélet modelljei alapján

Mint ahogy azt a reliabilitásmutatók elemzésekor is kifejtettük, az értékelők globális benyomását kifejező holisztikus szempont alapvetően másként viselkedik, mint az anali-

tikus skálák, amelyek valamilyen mértékben mind szerepet játszanak az általános ítélet kialakításában. Ezért a valószínűségi tesztelmélet módszereivel történő elemzésekbe is csak az analitikus szempontokat vontuk be. A vizsgálatból kizártuk a szövegek tartalmi és formai jegyeinek szétválasztása nélküli, a szöveg minőségéről alkotott általános képet reprezentáló összbenyomás szempontját.

A következőkben a valószínűségi tesztelmélet eszközeivel lefuttatott elemzések között a többaspektusú modellel és a parciális kreditmodellel történt számítások eredményeit és az ezekből levont következtetéseinket összegezzük. Bemutatjuk az értékelők szigorúsági paramétereit; a két bíráló ítéleteinek egymástól független elemzései alapján bemutatjuk az alkalmazott szempontok nehézségét és modellilleszkedését; majd elemezzük az értékelők skálahasználatában mutatkozó különbségeket.

Az értékelők szigorúsága

Az értékelők szigorúságának összehasonlítást a többaspektusú (*multifaceted*) modell teszi lehetővé. A vizsgálat során a modell a szempontokra adott ítéleteket elemezve felállítja az egyes bírálók szigorúsági paramétereit. Több bíráló esetén a legkevésbé és a leginkább szigorúan osztályozó értékelő, két bíráló esetén e kettő paramétereinek különbségét a szórással összevetve alkothatunk képet arról, milyen mértékű az értékelők szigorúsága közötti eltérés.

6. táblázat. Az értékelők szigorúsági paraméterei

Bírálók	*Szigorúsági paraméter*	*Infit paraméter*
Első	0,28	1,04
Második	-0,28	1,12

Az elemzés eredményei szerint (6. táblázat) a szövegkorpusz értékelését végző szakértők szigorúsági paraméterei között mindössze 0,56 logitegységnyi a különbség, ami a szórással (2,5) összehasonlítva alacsonynak tekinthető, vagyis az értékelők szigorúságában csak kicsi az eltérés.

Az értékelési szempontok nehézségi indexe és modellilleszkedése

Az értékelési szempontok nehézségi paramétereit a két értékelő esetén két külön parciáliskredit-modell felállításával vizsgáltuk, így az eredmények közvetetten hasonlíthatók össze. Nem az volt a célunk, hogy az egyes szempontokat összehasonlítsuk, hanem az, hogy az értékelők munkáját külön jellemezzük. A 7. és a 8. táblázatban közölt adatok alapján a szempontok nehézségét kifejező középérték-logitok az első bíráló értékelésében a helyesírás és a külalak kivételével az átlagos képességszint (0 logit) felettiek, a második bírálónál a tartalom, a nyelvhelyesség és a külalak nehézsége helyezkedik el az átlagos képességszint alatt. Mindkét bíráló a külalak és olvashatóság szempontot ítéli meg a legenyhébben és a szerkezet és kidolgozást a legszigorúbban.

7. táblázat. Az értékelési szempontok nehézségi indexe és modelllilleszkedése az első bíráló értékítéletei alapján

Szempontok	*Középérték-logitok*	*Standard hiba*	*Infit paraméter*
Tartalom	0,30	0,07	0,80
Feladattartás: szövegtípus	0,26	0,07	0,90
Feladattartás: hangnem	0,46	0,07	0,60
Szerkezet és kidolgozás	0,59	0,07	0,60
Stílus	0,18	0,07	0,80
Érthetőség	0,42	0,07	0,70
Nyelvhelyesség	0,03	0,07	0,80
Helyesírás, központozás	-0,12	0,07	2,00
Külalak és olvashatóság	-2,14	0,20	2,80

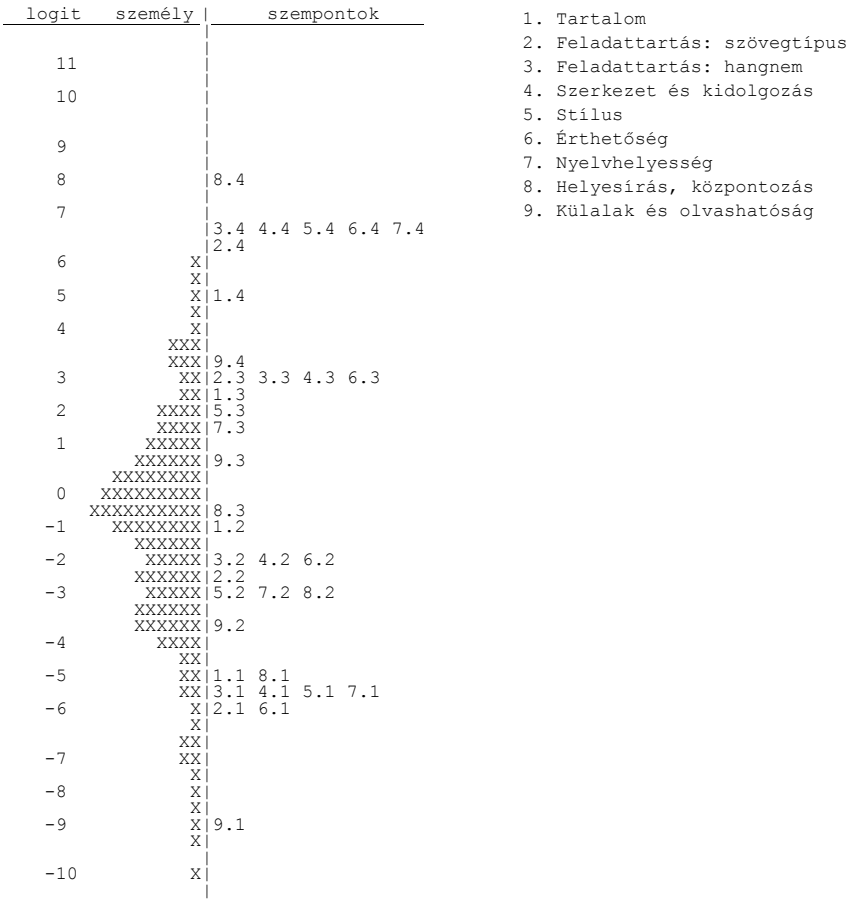
8. táblázat. Az értékelési szempontok nehézségi indexe és modelllilleszkedése a második bíráló értékítéletei alapján

Szempontok	*Középérték-logitok*	*Standard hiba*	*Infit paraméter*
Tartalom	-0,21	0,07	0,90
Feladattartás: szövegtípus	0,02	0,07	0,90
Feladattartás: hangnem	0,37	0,07	0,80
Szerkezet és kidolgozás	0,52	0,07	0,80
Stílus	0,38	0,07	0,80
Érthetőség	0,19	0,07	0,80
Nyelvhelyesség	-0,04	0,07	1,00
Helyesírás, központozás	0,07	0,07	1,60
Külalak és olvashatóság	-1,32	0,19	2,00

A szempontok modelllilleszkedését vizsgálva kutatásunkban *Park* (2004 idézi *Vígh*, 2010) nyomán az elfogadható értékek sávját 0,8 és 1,2 logit között határoztuk meg. Ennek alapján az első bíráló esetén a feladattartás: hangnem, a szerkezet és kidolgozás, az érthetőség, valamint a helyesírás, központozás és a külalak és olvashatóság sem illeszkedik jól a modellhez. A második bíráló esetén nem illeszkedő szempontként a helyesírás, központozást és a külalak és olvashatóságot kell kiemelnünk. E két szempont esetén az infit paraméter mindkét értékelőt tekintve 1,2 logit felett (2,0, és 2,8, illetve 1,6 és 2,0) helyezkedik el, vagyis ezeket a szempontokat többdimenzionalitás jellemzi, a többi szemponthoz képest mást mérnek, tőlük igen függetlenek, illetve értékelésükbe más szempontok is közrejátszanak (*Vígh*, 2010).

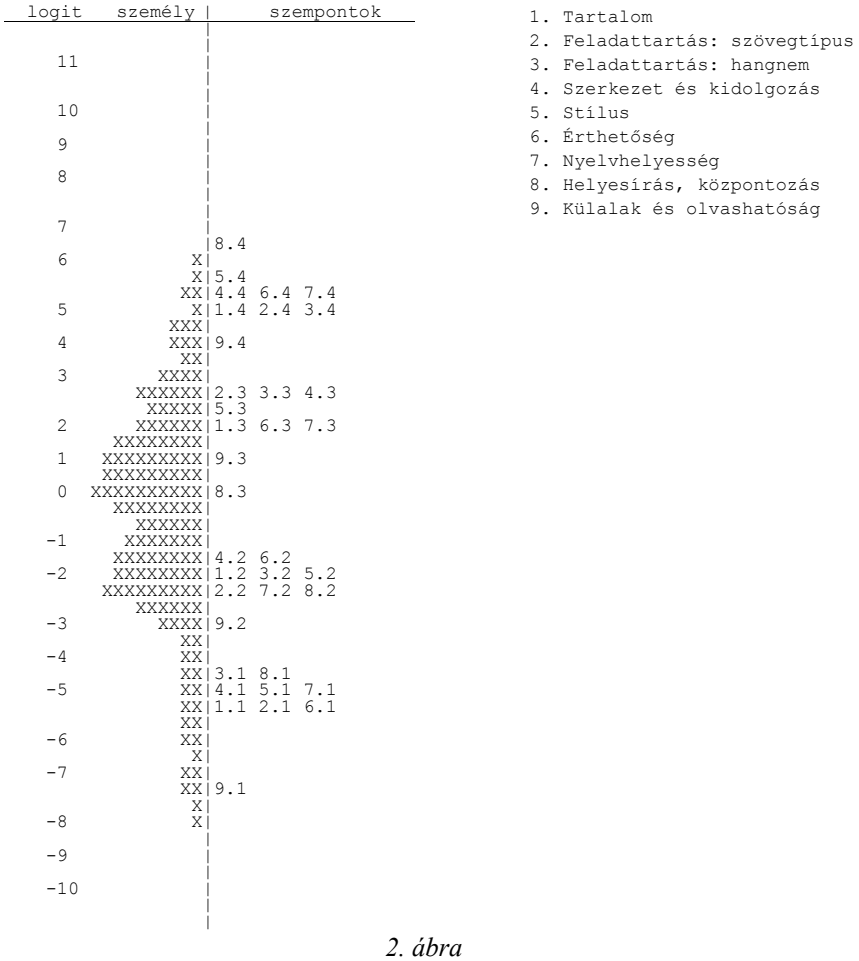
Az 1. és a 2. ábrán a két értékelő esetén külön lefuttatott parciáliskredit-modell outputjaként kirajzolt személy-szempont térképek láthatók, melyek közös skálán ábrázolják

a minta fogalmazásképességének fejlettségét és az értékelési skálapontok *Thrustone*-küszöbértékeit. Ez utóbbi értékek azokat a képességszinteket jelölik, amelyek mellett a tanulók 50%-os valószínűséggel kapják munkájukra adott szempontból az adott értékletet (*Molnár Gy*., 2008, 2013; *Vígh*, 2010). Az ábrák bal oldalán elhelyezkedő 11 és -10 közötti számsor (logitskála) értékei az átlaga alatti, átlagos és átlag feletti képességszinteket mutatják (*Molnár Gy*., 2008, 2013; *Vígh*, 2010). Az ábrák alapján a tanulók képességszintje nagyon széles skálán, 21 logitegységnyi területen helyezkedik el. A 90 fokban elforgatott eloszlásgörbére emlékeztető ábra X-ei mindkét ábrán körülbelül 3 tanulót jelölnek, akik a logitskála adott pontján helyezkednek el képességparamétereik alapján. Az ábrák jobb oldalán elhelyezkedő számok az egyes szempontok (1–9. magyarázatukat l. az 1. és a 2. ábra jobb oldalán) esetén az egyes skálapontok küszöbértékeit jelentik (*Molnár Gy*., 2008, 2013; *Vígh*, 2010).



1. ábra
Az értékelési szempontrendszer személy-szempont térképe az első bíráló értékítéletei alapján (X≈3 tanuló)

Az 1. és 2. ábráról leolvasható, hogy a szempontrendszer skálapontjai mennyire jól illeszkednek a vizsgált tanulók fogalmazásképességének fejlettségi szintjéhez. Az értékelési skálák legmagasabb foka nem működik megfelelően. A tanulók nagyon kis arányát jellemzik a legmagasabb értékletek, különösen az első bíráló ítéletei szerint. Az átlagos képességszint alatt mindkét bíráló esetén több tanuló helyezkedik el. A szempontrendszer működését jellemzi, hogy a rendszer a vizsgált mintához nem illeszkedik tökéletesen, nem fedi le kellő mértékben a tanulók képességszintjét, ugyanis a skála legmagasabb foka – különösen az első bíráló ítéletei szerint – csak a tanulók kis arányát jellemzi, a minta többsége a 2-es és a 3-as értékletet kifejező szintre került. Mindezek alapján a szempontrendszer a vizsgált mintánál jobb fogalmazásképességgel bíró tanulók esetén működik jól.



2. ábra
Az értékelési szempontrendszer személy-szempont térképe a második bíráló értékítéletei alapján (X≈3 tanuló)

Az értékelők skálahasználata

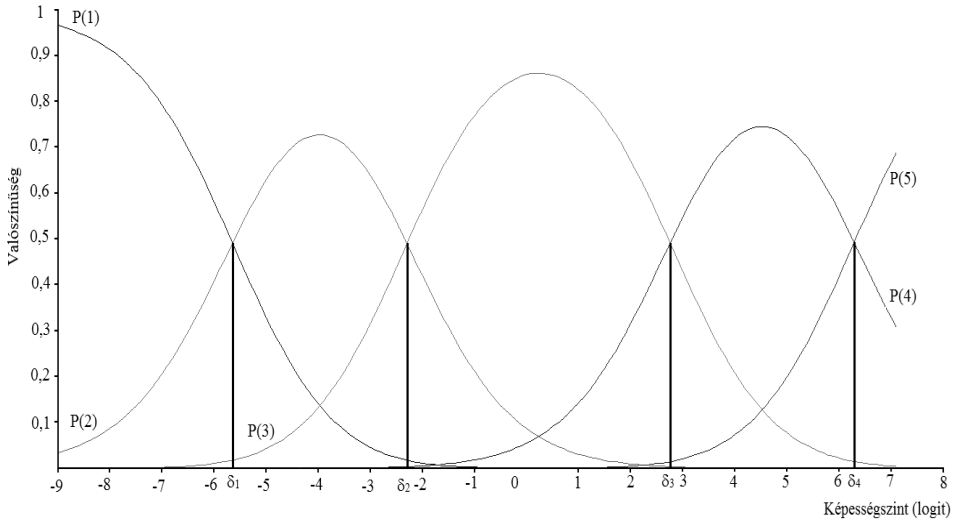
Az értékelők skálahasználatának jellemzését az itemek karakterisztikus görbéinek vizsgálata teszi lehetővé. A vízszintes tengelyen a képességszinteket, a függőlegesen a pontszámokhoz tartozás valószínűségét jelenítjük meg. A görbék az értékelési skálák egyes pontjainak karakterisztikus görbéi, annak valószínűségét jelenítik meg, hogy a különböző képességszintek mellett adott szempontból az adott skálapontra sorolják az értékelők a tanulót (*Molnár Gy*., 2008, 2013; *Vígh*, 2010).

Az értékelők skálahasználatának jellemzőiről további információkat adnak a görbék metszéspontjaiban elhelyezkedő δ-paraméterek. Ezek az értékek az egyes skálapontok közötti átmenetet jelölik; esetünkben azt mutatják meg, hogy a fogalmazásképesség fejlettségének mely logitértékben kifejezett szintjén azonos annak a valószínűsége, hogy két szomszédos skálapont egyikére sorolják a bírálók az adott tanulót. A δ-paraméterek két szempontból teszik lehetővé az értékelési skálák jellemzését. Távolságuk összehasonlításával megítélhetjük, milyen különbségek vannak az egyes skálapontok megszerzésének lehetőségében, hiszen minél nagyobb két δ-paraméter távolsága, annál nagyobb a valószínűsége annak, hogy az értékelők a tanulót az adott skálaponthoz rendelik. Másfelől érdemes vizsgálni a δ-paraméterek elrendeződését, vagyis azt, hogy nem cserélődnek-e fel a különböző értékek, ez ugyanis a bíráló skálahasználatának problémáit jelezné (*Molnár Gy*., 2008, 2013; *Vígh*, 2010).

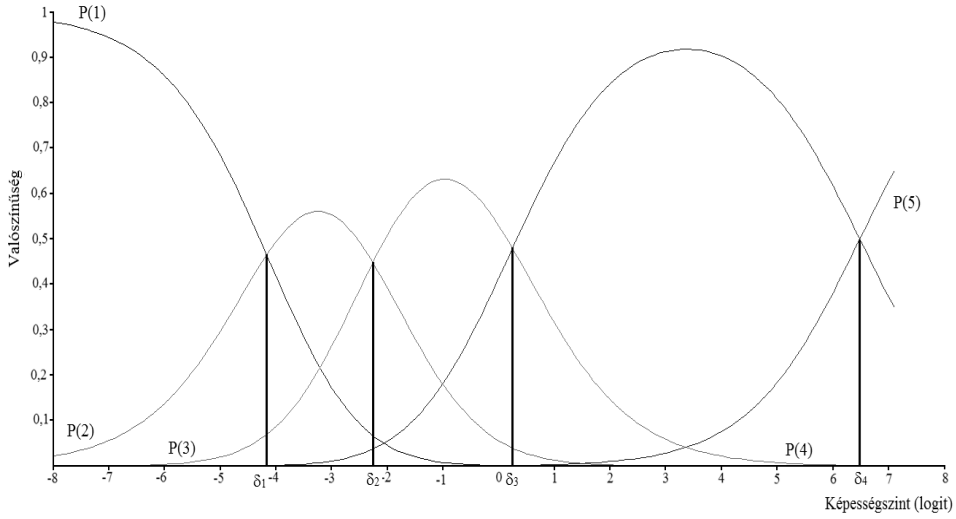
A következőkben csak néhány, jellemző működésű szempont görbéjét mutatjuk be, illetve jelezzük, melyek azok a szempontok, amelyek hasonlóan viselkednek. A 3. ábrán ideális skálaműködés látható. Amellett, hogy a δ-paraméterek megfelelően emelkednek, legnagyobb valószínűsége annak van, hogy a tanuló a hármas kategóriába kerül, közel azonos valószínűsége van a 2-es és a 4-es értéklet megszerzésének. Így működik az értékelési skála az első bíráló ítéletei alapján a feladattartás: szövegtípus mellett, a feladattartás: hangnem és az érthetőség szempontokon; valamint a második bíráló tartalom, szerkezet és kidolgozás, feladattartás: szövegtípus, feladattartás: hangnem, stílus, érthetőség, nyelvhelyesség, illetve külalak és olvashatóság szerint.

Az ideálistól eltérő, az értékelésben problémákat jelző itemkarakterisztikus görbékre ad példát a 4. és az 5. ábra. A 4. ábra alapján legnagyobb valószínűsége a 4-es értéklet megszerzésének van, ennél sokkal kisebb a 2-esé, és a 3-asé. A legtöbb tanuló tehát 4-es osztályzatot kap az így viselkedő szempontoknál, és alacsonnyá válik az ennél kisebb értékletet megszerzők aránya. Mindkét értékelő így használja például a helyesírásskálát, valamint az első bírálónál így működik a tartalom és a nyelvhelyesség szempont is.

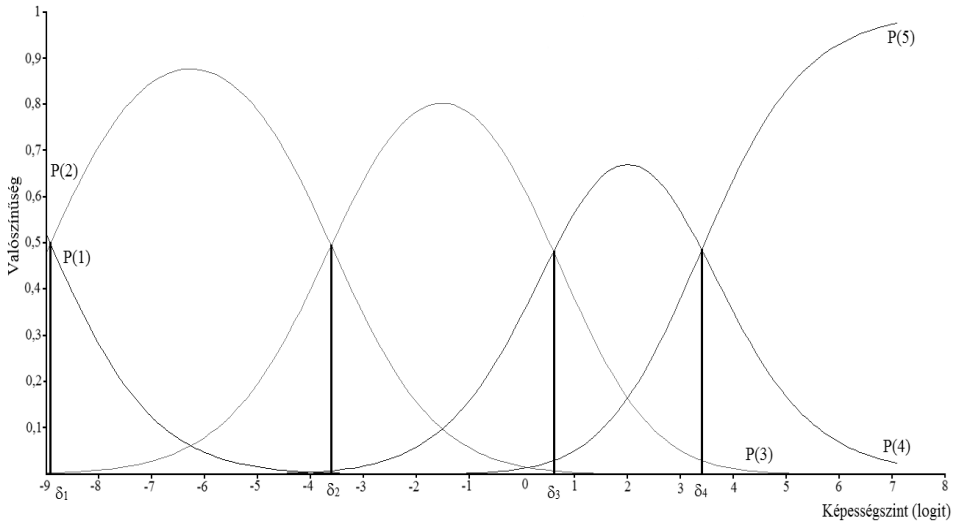
A rosszul működő értékelési skálára adnak példát az 5. ábrán megjelenített itemkarakterisztikus görbék. Megállapítható, hogy az első bíráló értékelésében a külalak és olvashatóság szempont esetén az 1-es skálapont gyakorlatilag nem működik, legnagyobb esélye a 2-esnek, valamivel kisebb a 3-asnak és még kisebb a 4-esnek van.



3. ábra
A feladattartás: szövegtípus szempont itemkarakterisztikus görbéi az első bíráló értékítéletei alapján



4. ábra
A helyesírás és központozás szempont itemkarakterisztikus görbéi a második bíráló értékítéletei alapján



5. ábra
A külalak és olvashatóság szempont itemkarakterisztikus görbéi az első bíráló értékítéletei alapján

Összegezve a valószínűségi tesztelmélet eszközeivel végzett vizsgálataink eredményeit: a parciáliskredit-modellel történő elemzéseink szerint a szigorúsági paraméterek kis eltérése ellenére sem tekinthetjük egyenértékűnek a két bíráló ítéleteit. Az értékelőket eltérő skálahasználat jellemezte, illetve az első bíráló esetén a szempontok modelllileszkedésében is problémákat azonosítottunk.

Összegzés és következtetések

Tanulmányunkban a fogalmazásképesség értékelésének kérdéseit vizsgáltuk. A szakirodalom összegzése alapján a szövegek minősítésének eszközeivel, módszereivel, valamint az értékelés megbízhatóságának problémáival foglalkoztunk. Ezt követően két független bíráló azonos tanulói szövegkorpusz minőségére adott ítéleteit a klasszikus és a valószínűségi tesztelmélet eszközeivel elemezve egy fogalmazásértékelési szempontrendszer működésére vonatkozóan fogalmaztunk meg következtetéseket. Empirikus vizsgálatunk célja az volt, hogy a korábbi, egy bíráló ítéleteivel dolgozó kutatásaink során megbízhatónak ítélt értékelési skálák működését két értékelő által adott osztályzatok további elemzésének alapján vizsgáljuk, a szempontrendszer használhatóságának érvényességére vonatkozóan árnyaltabb kijelentéseket fogalmazzunk meg. Hazai kontextusban kutatásunk újszerűségét az adta, hogy adatainkat a klasszikus tesztelmélet eszköztárába tartozó reliabilitásvizsgálatok, regresszióanalízisek és összefüggés-vizsgálatok mellett a valószí-

nűségi tesztelmélet módszereivel történő elemzésnek is alávetettük és ugyanazon tanulói szövegekre vonatkozóan rendelkeztünk két egymástól független értéklettel.

Mindkét bíráló esetén szignifikáns összefüggéseket tapasztaltunk az egyes szempontokra adott értékletek között, és a két értékelő azonos szempontokból megállapított minősítései szintén erős, szignifikáns kapcsolatban álltak egymással. Értékelőink szigorúságában nem találtunk nagy különbséget. Ennek ellenére jelentős eltéréseket mutattunk ki a bírálók értékelői teljesítményében. Míg a második értékelő munkáját jellemezve megállapítottuk, hogy az analitikus szempontok a szöveg külső megjelenítését és a helyesírási szabályok ismeretét kifejező osztályzatok kivételével kellő mértékben járultak hozzá az összbenyomás jegyéhez, illetve az említett két szempontot kivéve megfelelő modellilleszkedést mutattak, addig az első bíráló esetén számos problémát regisztráltunk. Láthattuk, hogy az ő értékelésében szignifikáns mértékben csak a tartalom, illetve a szerkezet és kidolgozás osztályzatok magyarázták a globális értéklet varianciáját. Több szempont esetén mutattunk ki az elfogadható intervallumon kívül elhelyezkedő infit paramétereket is, melyek a szövegértékelési skálák rossz modellilleszkedését jelzik.

Az értékelők skálahasználatának jellemzésére irányuló vizsgálataink alapján ugyancsak különbségeket találtunk a bírálók munkájában. A személy-szempont térképek alapján a szempontrendszer mintához való illeszkedése egyik bíráló esetén sem volt tökéletes, hiszen a legmagasabb képességszintre alig kerültek diákok. Ugyanakkor a szempontok itemkarakterisztikus görbéi és az egyes skálapontokra kerülés valószínűségét jelző δ-távolságok több esetben különböztek. Emellett az első értékelő munkájában találtunk rosszul működő szempontot is: a külalak és olvashatóság legalacsonyabb skálapontját gyakorlatilag nem használta a bíráló.

Ha az első bíráló munkáját tekintjük, a szempontrendszer működésében problémák mutatkoznak. A második bíráló értékletei alapján mind a klasszikus, mind a valószínűségi tesztelmélet eredményei az értékelési skálák megbízható működését igazolták. Szakirodalmi áttekintésünkben rámutattunk arra, hogy a bírálók személyiségbeli különbségeitől kezdve, a képzettségük eltérésein át számos oka lehet annak, hogy ugyanaz a szöveg a különböző bírálóktól eltérő értékelést kap. Vizsgálatunk során elképzelhető, hogy a szövegértékelés egyik, a szakirodalom elemzésekor is bemutatott problémájával, az egyes értékelési szempontok, illetve a hozzájuk rendelt skálapont-definíciók eltérő értelmezésének jelenségével szembesültünk. Az értékelésekben kimutatott különbségek esetünkben adódhattak a szempontrendszer és az értékelők képzésének elégtelenségéből egyaránt. Így szükségesnek látszik mind az értékelési skálák felülvizsgálata, mind további szövegvizsgálatok esetén az értékelők felkészítésének több részletre kiterjedő lebonyolítása is.

Az alkalmazott értékelési szempontrendszer továbbfejlesztése, illetve a későbbi fogalmazáskutatások módszereinek és eszközeinek megválasztása, kidolgozása szempontjából is fontos kiemelnünk, hogy mind a klasszikus, mind a valószínűségi tesztelmélet eszközeivel végzett elemzéseink rávilágítottak arra, hogy a helyesírás, központozás, valamint a külalak és olvashatóság osztályzatok viszonylag függetlennek bizonyultak a többi szemponttól (*Molnár E. K.*, 2000). A két bíráló által adott osztályzatok összefüggéseit kutató korrelációszámítások és a regressziós modellek eredményei azt mutatták, hogy ezek azok a szempontok, amelyek legkisebb mértékben korrelálnak a fogalmazás-

képesség egyéb összetevőinek fejlettségét kifejező szempontokon elért teljesítményekkel, illetve a legkevésbé járulnak hozzá az értékelők szövegekről alkotott globális ítéleteinek varianciájához. A helyesírás, központozás, illetve a külalak és olvashatóság szempontok függetlenségét igazolták a parciáliskredit-modellel lefuttatott számítások is. A szempontok modellilleszkedését jelző infit paraméterek mindkét bíráló esetén az elfogadható intervallum feletti értékeket vettek fel e két skála esetén, ami szintén a helyesírási jellemzők és a szöveg külső megjelenítésének egyéb szempontoktól való függetlenségét, többdimenzionalitását jelzi. Mindezek alapján megfontolandó a helyesírás és a külalak más, a többi szövegjellemzőt vizsgáló skálától független szempontrendszerrel történő értékelése és ezektől történő elválasztása például a *Vidákovich* (1986), illetve *Orosz* és *Vidákovich* (1988) szempontjai szerint.

További kutatási feladat annak feltárása, milyen egyéb okok húzódnak meg az értékelések különbözősége mögött, befolyásoló tényező-e az értékelők korábbi mérés-értékelési, illetve tanítási tapasztalata. Ennek megállapításához újabb értékelők bevonása, a skálapontok definícióinak közös értelmezése lehet szükséges. A szempontrendszer továbbfejlesztése mind a későbbi fogalmazáskutatások, mind a pedagógusok fogalmazás-értékelési gyakorlatának fejlesztéséhez fontos lehet. Egy kezelhető számú szemponttal, jól értelmezhető skálapontokkal rendelkező, a gyakorlatban megfelelően használható értékelési rendszer kiindulópontul szolgálhat a többszempontú iskolai fogalmazásértékelési kultúra meghonosításához, ezáltal a tanulók fogalmazásképességének fejlesztéséhez is.

Köszönetnyilvánítás

Köszönjük *Molnár Gyöngyvér* és *Vígh Tibor* adatelemzésben nyújtott segítségét. A tanulmányban közölt empirikus eredményeket a X. Pedagógiai Értékelési Konferencián elhangzott előadásban (*Nagy*, 2012) bemutattuk.

A kutatás megvalósítását a TÁMOP 3.1.9-08/1-2009-0001 és a TÁMOP 3.1.9-11/1-2012-0001 támogatta.

Irodalom

Barkaoui, K. (2007): Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing*, **12**. 2. sz. 86–107.

Barkaoui, K. (2011): Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles*, *Policy and Practice*, **18**. 2. sz. 279–293.

Beaugrande, R. A. D. (1984): *Text production: Toward a science of text production*. Ablex, Norwood.

Bereiter, C. (1980): Development in writing. In: Gregg, L. W. és Steinberg, E. R. (szerk.): *Cognitive processes in writing*. L. Erlbaum Associates, Hillsdale. 73–93.

Bereiter, C. és Scardamalia, M. (1987a): Knowledge telling and knowledge transforming in written composition. In: Rosenberg, S. (szerk.): *Advances in applied linguistics*. *Vol. 2. Reading, writing and language learning*. Cambridge University Press, Cambridge. 142–175.

Bereiter, C. és Scardamalia, M. (1987b): *The psychology of written composition*. Lawrence Erlbaurn Associates, Publishers, Hissdale, New Jersey London.

Beyreli, L. és Ari, G. (2009): The use of analytic rubric in the assessment of writing performance – inter-rater concordance study. *Educational Sciences: Theory and Practice*, **9**. 1. sz. 105–125.

Chai, C. (2006): Writing plan quality: Relevance to writing scores. *Assessing Writing*, **11**. 198–223.

Crawford, L. és Smolkowski, K. (2008): When a „sloppy copy" is good enough: Results of a state writing assessment. *Assessing Writing*, **13**. 1. sz. 61–77.

Dávid Gergely (2008): Az emelt szintű idegen nyelvi érettségi és az államilag elismert nyelvvizsgák a vizsgázói teljesítmények tükrében. Összegző tanulmány, bővített változat. Nyelvvizsgáztatási Akkreditációs Központ. Budapest.

Davis, W. (2005): The effects of grammar testing on the writing quality and reduction of errors in college freshmen's essays. http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/1b/c1/a7.pdf. Utolsó letöltés: 2010. április 7.

Eckes, T. (2008): Rater types in writing performance assessments: a classification approach to rater variability. *Language Testing*, **25**. 2. sz. 155–185.

Eckes, T. (2005): Examining rater effects in TestDaF writing and speaking performance assessments: a many-facet Rasch analysis. *Language Assessment Quarterly*, **2**. 3. sz. 197–221.

Engelhard, G. és Myford, C. M. (2003): *Monitoring faculty consultant performance in the Advanced Placement English Literature and Composition Program with a many-faceted Rasch model*. College Entrance Examination Board, New York.

Engelhard, G. Jr. (1994): Examining rater errors in the assessment of written composition with a many faceted Rasch model. *Journal of Educational Measurement*, **31**. 2. sz. 93–112.

Engelhard, G. Jr., Gordon, B. és Gabrielson, S. (1991): Writing task and the quality of student writing: Evidence from a statewide assessment of writing. Előadás, Annual Meeting of the American Educational Research Association, Chicago. 1991. április 3–7. http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/22/ef/b7.pdf. Utolsó letöltés: 2010. április 7.

Eysenck, M. W. és Keane, M. T. (1997): *Kognitív pszichológia*. Nemzeti Tankönyvkiadó, Budapest.

Flower, L. és Hayes, J. (1980): The dynamics of composing: making plans and juggling constraints. In: Gregg, L. W. és Steinberg, E. R. (szerk.): *Cognitive processes in writing*. L. Erlbaum Associates, Hillsdale. 31–50.

Gearhart, M., Herman, J. L., Novak, J. R. és Wolf, S. A. (1995): Toward the instructional utility of large-scale writing assessment: Validation of a new narrative rubric. *Assessing Writing*, **2**. 2. sz. 207–242.

Gelati, C. és Boscolo, P. (2009): Improving the quality of primary school children's narration of personal events. An intervention study on the use of evaluation strategies. *L1 – Educational Studies in Language and Literature*, **9**. 3. sz. 1–28.

Gorman, T. P., Purves, A. C. és Degenhart, R. E. (1988, szerk.): *The IEA study of written composition I. The international writing tasks*. Pergamon Press, Oxford.

Griffin, P. és Anh, P. N. (2005): Assessment of creative writing in Vietnamese primary education. *Asia Pacific Education Review*, **6**. 1. sz. 72–86.

Gyagenda, I. S. és Engelhard, G. (1998): Applying the Rasch model to explore rater influences on the assessed quality of students' writing ability. Előadás, Annual Meeting of the American Educational Research

Association, San Diego. 1998. április 13–17. http://www.eric.ed.gov/PDFS/ED422350.pdf. Utolsó letöltés: 2012. január 20.

Hayes, J. és Flower, L. (1980): Identifying the organization of writing processes. In: Gregg, L. W. és Steinberg, E. R. (szerk.): *Cognitive processes in writing*. L. Erlbaum Associates, Hillsdale. 3–30.

Hayes, J. R. (1996): A new framework for understanding cognition and affect in writing. In: Levy, C. M. és Ransdell, S. (szerk.): *The science of writing: Theories*, *methods*, *individual differences and applications*. Lawrence Erlbaum Associates, Mahwah, N. J. 1–27.

Hidi, S. és Boscolo, P. (2007, szerk.): *Writing and motivation*. Elsevier, Amesterdam.

Hillocks, G. (1986): *Research on written composition*. Clearinghouse on Reading and Communication Skills, Urbana, Illinois.

Horváth Zsuzsanna (1998): *Anyanyelvi tudástérkép. Középiskolai tantárgyi feladatbankok III*. Országos Közoktatási Intézet, Budapest.

Isonio, S. (1991): Judgments of placement writingsampels at golden west collage: an evaluation of inter-rater reliability. Kézirat. Golden West College, Huntington Beach. http://www.eric.ed.gov/PDFS/ED345784.pdf. Utolsó letöltés: 2012. február 3.

Kádárné Fülöp Judit (1990): *Hogyan írnak a tizenévesek? – Az IEA-fogalmazásvizsgálat Magyarországon*. Akadémiai Kiadó, Budapest.

Kontra József (2009): A parciális kredit modell egy alkalmazása. In: Psenáková Ildikó, Mező Ferenc és Viczayová Ildikó (szerk.): *Képzés és gyakorlat II*. Konstantin Filozófus Egyetem, Nyitra. 99–108.

MacArthur, C. A., Graham, S. és Fitzgearld, J. (2008, szerk.): *Handbook of writing research*. The Guilford Press, New York.

Mäki, H. (2002): *Elements of spelling and composition studies on predicting and supporting writing skills in primary grades.* Turun Yliposto, Turku.

Molnár Edit Katalin (1996): A kognitív pszichológia három fogalmazás-modellje. *Magyar Pedagógia*, **96**. 2. sz. 139–156.

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

Molnár Edit Katalin (2002): Az írásbeli szövegalkotás. In: Csapó Benő (szerk.): *Iskolai műveltség*. Osiris, Budapest. 193–216.

Molnár Edit Katalin (2003): Az írásbeli szövegalkotás fejlődése: vizsgálatok 10–17 éves tanulók körében. Doktori disszertáció. ELTE BTK Neveléstudományi Doktori Iskola, Budapest.

Molnár Edit Katalin (2009): Az írásbeli szövegalkotás funkciója és hatékonysága magyar egyetemista diákok dolgozatainak szövegeiben. *Anyanyelv-pedagógia*, **2**. 1. sz. http://www.anyanyelv-pedagogia.hu/cikkek.php?id=138. Utolsó letöltés: 2013. szeptember 6.

Molnár Gyöngyvér (2003): Az ismeretek alkalmazásának vizsgálata modern tesztelméleti (IRT) eszközökkel. *Magyar Pedagógia*, **103**. 4. sz. 423–446.

Molnár Gyöngyvér (2005): Az objektív mérés lehetősége: a Rasch-modell. *Iskolakultúra*, **15**. 3. sz. 71–80.

Molnár Gyöngyvér (2006): A Rasch-modell alkalmazása a társadalomtudományi kutatásokban. *Iskolakultúra*, **16**. 12. sz. 99–113.

Molnár Gyöngyvér (2008): A Rasch-modell kiterjesztése nem dichotóm adatok elemzésére: a rangskálás és a parciális kredit modell. *Iskolakultúra*, **18**. 1–2. sz. 66–77.

Molnár Gyöngyvér (2013): *A Rasch modell alkalmazási lehetőségei az empirikus kutatások gyakorlatában. Alapvető elemzések a társadalomtudományi kutatásokban*. Gondolat Kiadó, Budapest.

Molnár Gyöngyvér és Józsa Krisztián (2006): Az olvasási képesség értékelésének tesztelméleti megközelítései. In: Józsa Krisztián (szerk.): *Az olvasási képesség fejlődése és fejlesztése.* Dinasztia Tankönyvkiadó, Budapest. 155–174.

Nagy József (1996): *Nevelési kézikönyv: Személyiségfejlesztő pedagógiai programok készítéséhez*. Mozaik Oktatási Stúdió, Szeged.

Nagy József (2002): *XXI. század és nevelés*. Osiris Kiadó, Budapest.

Nagy Zsuzsanna (2009): 17 éves tanulók szövegalkotási képessége és szövegekre vonatkozó ítéletei. *Iskolakultúra*, **19**. 11. sz. 19–31.

Nagy Zsuzsanna (2010): A fogalmazásképesség értékelésének lehetősége egy próbamérés eredményei alapján. Előadás, Országos Neveléstudományi Konferencia, Budapest. 2010. november 4–6. 186.

Nagy Zsuzsanna (2011): Magyart tanító pedagógusok fogalmazástanítással kapcsolatos meggyőződései és oktatási módszerei. Előadás, XI. Országos Neveléstudományi Konferencia, Budapest. 2011. november 3–5. 171.

Nagy Zsuzsanna (2012): Fogalmazásértékelési szempontrendszer működésének vizsgálata két független bíráló ítéletei alapján. Előadás, X. Pedagógiai Értékelési Konferencia, Szeged. 2012. április 26–28. 102.

Orosz Sándor (1972): *A fogalmazástechnika mérésmetodikai problémája és országos színvonala*. Tankönyvkiadó, Budapest.

Orosz Sándor és Vidákovich Tibor (1988): Anyanyelvi helyesírásvizsgálatok. Eredménymérések néhány módszertani kérdése. *Pedagógiai Technológia*, 2. sz. 30–40.

Park, T. (2004): An investigation of an ESL placement test of writing using many-facet Rasch measurement. *Working Papers in TESOL and Applied Linguistics*, **4**. 1. sz. 1–21. o.

Pintér Henrett (2009): Erkölcsi gondolkodás 9–10 évesek írásbeli szövegalkotásában. *Iskolakultúra*, **19**. 10. sz. 109–116.

Popp, S. E. O., Ryan, J. M., Thompson, M. S. és Behrens, J. T. (2003): Operationalizing the rubric: The effect of benchmark selection on the assessed quality of writing. Előadás, Annual Meeting of the American Educational Research Association, Chicago. 2003. április 1–25. http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/1b/7c/cf.pdf. Utolsó letöltés: 2010. április 7.

Purves, A. (1992): Reflections on research and assessment in written composition. *Research in the Teaching of English*, **26**. 1. sz. 108–22.

Ransdell, S. E. és Levy, C. M. (1996).Working memory constraints on writing performance. In: Levy, C. M. és Ransdell, S. E. (szerk.): *The science of writing*. Lawrence Erlbaum Associates, Mahwah, New Jersey. 93–101.

Ransdell, S., Levy, C. M. és Kellogg, R. (2002): The structure of writing processes as revealed by secondary task demands. *L1 – Educational studies in Language and Literature*, **2**. 2. sz. 141–163.

Schoonen, R. (2005): Generalizability of writing scores: An application of structural equation modeling. *Language Testing*, **22**. 1. sz. 1–30.

Segev-Miller, R. (2004): Writing from sources: The effect of explicit instruction on college students' processes and products. *L1 – Educational Studies in Language and Literature*, **4**. sz. 5–33.

Sudweeks, R. R., Reeve, S. és Bradshaw, W. S. (2004): A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing*, **9**. 3. sz. 239–261.

Sugita, Y. (2009): The development and implementation of task-based writing performance assessment for Japanese learners of English. *Journal of Pan-Pacific Association of Applied Linguistics*, **13**. 2. sz. 77–103.

Szilassy Eszter (2012): Az írás és fogalmazásjavítás stratégiái. *Anyanyelv-pedagógia*, **4**. 1. sz. http://www.anyanyelv-pedagogia.hu/cikkek.php?id=357. Utolsó letöltés: 2013. szeptember 6.

Takala, S. (1988): Origins of the international study of writing. In: Gorman, T. P., Purves, A. C. és Degenhart, R. E. (szerk.): *The IEA study of written composition I.: The international writing tasks and scoring scales*. Pergamon, Oxford. 3–14.

Torrance, M., van Waes, L. és Galbraith, D. (2007, szerk.): *Writing and cognition: research and application*. Elsevier, Amsterdam.

Vidákovch Tibor (1990): *Diagnosztikus pedagógiai értékelés.* Akadémiai Kiadó. Budapest.

Vidákovich Tibor (1986): Az íráskészség vizsgálatának néhány mérésmetodikai problémája. *Acta Universitatis Szegediensis de Attila József Nominatae Sectio Paedagogica et Psychologica*, **28**. 117–138.

Vígh Tibor (2008): Egy IRT-alapú nyelvi feladatban létrehozásának módszertani kérdései. A német érettségi vizsgafeladatok elemzésének eredményei. *Magyar Pedagógia*, **108**. 1. sz. 29–51.

Vígh Tibor (2010): Az idegen nyelvi érettségi működés és hatása a tanulói teljesítmények és a tanári nézetek tükrében. PhD-értekezés. Szegedi Tudományegyetem, Neveléstudományi Doktori Iskola, Szeged.

Weigle, S. C. (1998): Using FACETS to model rater training effects. *Language Testing*, **15**. 2. sz. 263–287.

Whithaus, C., Harrison, S. B. és Midyette, J. (2008): Keyboarding compared with handwriting on a high-stakes writing assessment: Student choice of composing medium, raters' perceptions, and text quality. *Assessing Writing*, **13**. 5. sz. 4–25.

Wiseman, C. S. (2012): Rater effects: ego engagement in rater decision-making *Assessing Writing*, **17**. 3. sz. 150–173.

Wu, M., Adams, R. J. és Wilson, M. R. (1998): *ACER ConQuest. Generalised Item Response Modelling Software*. ACER Press, Australia.

Zhang, L. és Vukelich, C. (1998): Prewriting activities and gender: Influences of the writing quality of male and female students. Előadás, Annual Meeting of the American Educational Research Association, San Diego. 1998. április 13–17.
http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/15/b4/5a.pdf. Utolsó letöltés: 2010. április 7.

ABSTRACT

ZSUZSANNA NAGY: THE RELIABILITY OF WRITTEN COMPOSITION ASSESSMENT BASED ON AN ANALYSIS OF THE EVALUATIONS OF TWO INDEPENDENT RATERS

The purpose of the present study is to examine the operation of our scoring system, which was developed to assess the quality of schoolchildren's texts and to analyse two independent raters' evaluations. Rater performance was analysed using classical test theory and item response theory. The sample included 429 Hungarian children in Year 8. They were asked to produce a narrative text, and their compositions were scored by two independent raters using our scoring system, which comprises one holistic criterion and nine analytic ones (content, genre, tone, organization and structure, style, readability, lexical and grammatical conventions, spelling and orthographic conventions, handwriting and neatness). The analyses showed high reliability (Cronbach-α=0.95) for both raters. There are strong and significant correlations between the ratings (r=.85-.93, p<.01) and small differences (.56 logit) between the severity parameters of the two raters. The partial credit model analyses revealed differing uses of the scales by the raters. The scores given by the first rater on most scales show a bad model fit. The δ parameters of the scales' characteristic curves indicate that the two raters used the scales differently. Results call attention to the problem of defining scales for written composition scoring systems. The results point out that a precise definition of the scales was unable to guarantee objective and consistent assessment as the two raters still interpreted the scales differently. The findings indicate the need to re-examine the scoring system used and to provide training for raters.

Magyar Pedagógia, **113**. Number 3. 153–179. (2013)

Levelezési cím / Address for correspondence: Nagy Zsuzsanna, SZTE Oktatáselméleti Kutatócsoport, H–6722 Szeged, Petőfi S. sgt. 30–34.