A javító és az értékelőskála interakciójának vizsgálata az írásfeladatok értékelésének validálása során

1 A disszertáció témája és célja

A vizsgálat célja egy használatban levő értékelőskála validálása, amelyet középfokú írásfeladatok értékelése során alkalmazunk. Két különböző kutatási módszerrel kívántuk a szubjektív értékelésű írásfeladatok javítása során fellépő mérési hibát azonosítani. Egyrészről, modern tesztelméleti eszközök segítségével, kvantitatív módszerekkel vetettük vizsgálat alá a hatfokozatú értékelőskálát. Vizsgálódásunkat tovább bővítettük a vizsgáztatói magatartás megfigyelésével: az értékelési folyamat komplexitásáról hangos gondolkodás módszerrel és interjúkkal gyűjtött adatok segítségével kívántunk bővebb információt nyerni.

Az értékelési folyamat validitását az értékelők és az értékelőskála interakciójának vizsgálata alapján tudjuk megerősíteni. A kapott eredmények mind az értékelők, mind az értékelőskála pszichometriai validitását alátámasztják. Apróbb problémák és mérési hiba lehetséges forrásai ugyanakkor nyilvánvalóvá váltak.

A kutatás felveti annak lehetőségét, hogy az alkalmazott modern tesztelméleti módszereket szélesebb körben is használjuk további szubjektív értékelésű feladatok értékelési folyamatainak validálása során, melyhez e projekt megfelelő kiindulási alapul szolgálhat. A vizsgálat gyakorlati eredményei továbbá beépíthetők a napi tesztfejlesztési munkálatokba azzal a céllal, hogy az adott körülmények között a legmegbízhatóbb és érvényesebb mérési eljárásokat tudjuk alkalmazni.

E fent körvonalazott kutatást két általános és hat specifikus kutatási kérdés irányította. Az első kérdéskör az értékelési folyamat során tapasztalható nemkívánatos interakciós mintákat volt hivatott feltárni, míg a második kérdéskör, amelynek középpontjában az értékelői viselkedés áll, ezek okaira kívánt rávilágítani.

A következő kérdések segítségével az értékelőskálát vizsgáltuk:

I. Mivel igazoljuk a mérőeszközünk pszichometriai validitását?

1. Mely értékelési szempontok váltanak ki elfogultságot az értékelőből?
2. Melyik értékelési kritériumok esetében tapasztalható variancia hiánya?
3. Milyen mértékben érzékelhető a holdudvar hatás, vagyis egy értékelési szempont domináns szerepe?
4. Milyen mértékben igazolják az eredmények az értékelőskála hat sávjának és négy szempontjának megfelelő működését?

A következő kérdések az értékelő és az értékelési szempontok interakcióját vizsgálták:

1. Milyen tényezők váltanak ki a vizsgáztatókból olyan reakciókat, amelyek eltérítik őket az értékelőskála szerinti érékeléstől?
2. Milyen tényezők merülnek fel a javítás során, amelyek nem függnek össze a mérendő tulajdonsággal?

2 A dolgozat felépítése

Az *első fejezet* bemutatja azt a kutatási hátteret, amelyeket a következő fejezetek részletesen is kifejtenek. A vizsgálat elméleti hátterének rövid bemutatása mellett a fejezet felvázolja a gyakorlati célokat is. Ebben a részben felsoroljuk a két tágabb és a hat specifikus kutatási kérdést is.

A *második fejezetben* bemutatásra kerül az az elméleti háttér, amely kontextust szolgáltat a kutatásnak. Elsőnek a leírás néhány, a nyelvi tesztek validálására vonatkozó alaptétel tárgyalását tartalmazza. A fejezet második része a mérési hibával foglalkozik: a mérési hiba általános bemutatásától elindulva a leírás bemutatja a mérési hiba azon típusait, amelyek a szubjektív értékelésű feladatokra jellemzőek. Ezt követi azoknak a tesztfejlesztés

során is alkalmazott validálási eljárásoknak a bemutatása, amelyek a modern tesztelméletre épülnek. Végül szó esik a legfontosabb teoretikus művekről, amelyek az íráskészség értékelését modellezik.

A harmadik fejezet bemutatja a kutatás empirikus hátterét. Áttekintést ad azokról a legfontosabb kutatásokat, amelyek az értékelői variabilitással foglalkoznak. E kutatás fókusza az értékelő és az értékelőskála interakciójának vizsgálata, ugyanakkor az értékelési folyamat más tényezői is közvetett hatással lehetnek erre a kapcsolatra. Az empirikus háttér bemutatását a tesztfejlesztésben gyakran alkalmazott validálási módszerek összefoglalása vezeti be. Ezt követi a releváns irodalom ismertetése, különös figyelemmel az értékelő és az értékelőskála közötti kapcsolat vizsgálatát tárgyaló tanulmányokra. A fejezetet a modern tesztelmélet magyarországi alkalmazásának bemutatása zárja.

A kutatási módszerek bemutatására a *negyedik fejezetben* kerül sor. Ezeket a két külön tanulmánynak megfelelően (Study 1, Study 2) külön részekben tárgyaljuk. Az adatgyűjtés részletes leírását mind a kvantitatív első, és a kvalitatív második tanulmányra vonatkozóan hasonló megosztásban az adatelemzés leírása követi. Az elemzési módszerként alkalmazott többitemparaméteres Rasch modell által kínált eredményeknek csak egy részét fogjuk felhasználni; ezek rövid bemutatása és értelmezése ebben a fejezetben kapott helyet.

Az *ötödik fejezet* részletesen tárgyalja az első tanulmány eredményeit. A többitemparaméteres Rasch analízis segítségével képet kapunk arról, hogy az értékelők hogyan értelmezik az értékelőskálát, és mennyire következetesek a hatfokú analitikus értékelőeszköz használatában. Eredményeink megmutatják, hogy esetlegesen hol merülnek fel problémák az értékelősávok és az értékelési szempontok használata során.

Az értékelői magatartás vizsgálatának eredményeit a *hatodik és a hetedik fejezet* tartalmazza. Az előző fejezetben a kvantitatív adatok segítségével feltárt értékelői tulajdonságok itt kiegészülnek további adatokkal: a hangos gondolkodás módszer segítségével gyűjtött adatok (hatodik fejezet), továbbá az interjúkból származó információk (hetedik fejezet) fontos kiegészítésekkel szolgálnak az értékelési folyamat minden tényezőjéről. Az adatok nem csupán az értékelőre és az értékelési szempontokra vonatkoznak, hanem kapcsolatot teremtenek a feladat, a teljesítmény, a vizsgázó, az értékelő, a pontszám és az értékelési szempontok között. A fejezet végén összevetjük a különböző forrásokból származó adatokat.

A dolgozatot záró *nyolcadik fejezetben* választ adunk a kutatási kérdésekre, és eredményeinket összevetjük korábbi, hasonló kutatások eredményeivel. Ebben a fejezetben tárgyaljuk vizsgálatunk eredményeinek implikációit, és felvetjük az eredmények általánosíthatóságának kérdését más szubjektív értékelésű feladatokra vonatkozóan. Felsoroljuk továbbá azokat a problémás pontokat, amelyek kiküszöbölésére a kutatás során nem volt módunk, de amelyeket eredményeink értelmezése során figyelembe vettünk.

3 Elméleti háttér

Mivel a kutatás célja egy mérőeszköz validitásának megállapítása a mérési hiba feltárása segítségével az íráskészség értékelése során, az elméleti háttér bemutatása három területet ölel föl. Elsőként a klasszikus tesztelméletben alkalmazott mérési hiba fogalmával foglalkozunk (Spearman, 1904; Crocker & Algina, 1986), majd rámutatunk azokra a pontokra, amelyekben a modern tesztelmélet (Carmines & Zeller 1979; Crocker & Algina, 1986; Nunnally & Bernstein, 1994; Viswanathan, 2005) előrelépést mutat a klasszikus tesztelmélet mérési hiba felfogásához képest.

Annak ellenére, hogy a klasszikus tesztelmélet a legtöbb mérési modell alapja, és több kísérlet is történet arra, hogy a szubjektív értékelésű feladatok értékelésére is alkalmazható legyen (Choppin, 1982; De Gruiter, 1984; Saal, Downey & Lahey, 1980), a többitemparaméteres Rasch modell (Linacre, 1989) teszi lehetővé azt, hogy a szubjektív értékelésű feladatokat is nagyfokú objektivitással tudjuk értékelni. Ezért, a szubjektív értékelőskála validálásában a második fontos elméleti alap, amely segítségével a kutatás építkezik, a többitemparaméteres Rasch modell. Ez a teória az egyparaméteres Rasch modell kiterjesztése, amely a feladat nehézségén és a vizsgázó képességén túl további paraméterek modellezésére is lehetőséget ad. A módszer különösen jól alkalmazható olyan esetekben, amikor az értékelésben a szubjektív emberi tényező is szerepet játszik, mivel lehetőséget kínál arra, hogy ezeket a szubjektív elemeket azonosítsa. A szubjektív értékelésű feladatok értékelése során tipikusan a következő tényezők játszanak fontos szerepet: vizsgázó képessége, feladat nehézsége, értékelő szigorúsága és az értékelőskála tulajdonságai. Az értékelői hiba (Saal, Downey & Lahey, 1980, Engelhard, 1994; Linacre, 2003-6; Wolfe, Moulder, Bradley & Myford, 2001) áll a jelen vizsgálódások középpontjában, mivel ennek jelenléte nagymértékben veszélyeztetheti az értékelés megbízhatóságát és validitását.

A harmadik terület, amely teoretikus modelljeit fontos számba vennünk, az íráskészség értékelése (Engelhard, 1992; Lumley, 2005). Ezek a modellek azonosítják azokat a változókat, amelyek mivel kapcsolatot teremtenek a mért készség és az arra adott megfigyelt pontszám között, az értékelési folyamat során arra hatással vannak és azt befolyásolják. Ezek a dimenziók olyan lehetséges értékelői variabilitást és mérési hibát rejthetnek, amelyeket a többitemparaméteres Rasch modell segítségével fel tudunk tárni.

4 Kutatási módszerek

A kutatási kérdésekre két lépcsőben, kvantitatív (Study 1) és kvalitatív (Study 2) módszerek segítségével kerestünk választ.

4.1 Adatgyűjtés és –elemzés az első tanulmányban

Összesen három év alatt, hét alkalommal gyűjtött 2011 dolgozatot vizsgáltunk, melyet 27 értékelő javított. Az ezekre a dolgozatokra adott analitikus pontszámok biztosították az első tanulmány adatbázisát. A többitemparaméteres Rasch analízist a FACETS (Version 3.61.0) szoftverrel végeztük. Az elemzéshez a hatfokú, négy kritériumból álló értékelőskála alapján adott pontszámokat vizsgáltuk, amelyeket az értékelők a kettős javítás első fázisában adtak. Úgy gondoltuk, hogy ezek az elsődleges pontszámok pontosabb képet adnak a vizsgáztatói magatartásról, mint az egyeztetett, végső pontszámok. Az első tanulmányban a következő kérdésekre kerestünk választ.

1. Mely értékelési szempontok váltanak ki elfogultságot az értékelőből?
Az elemzés eredményében a standardizált t vagy z értékeket vizsgáljuk. Amennyiben ezekre 2-nél nagyobb értéket kapunk, akkor szignifikáns interakciós hatást tapasztalhatunk az értékelő és az értékelőskála között. Fontos ugyanakkor megjegyezni, hogy csak a visszatérően tapasztalt interakció hatást tarthatjuk valódi elfogultságnak.

2. Mely értékelési kritériumok esetében tapasztalható variancia hiánya?
Erre a kérdésre választ az illeszkedési statisztika vizsgálatával kapunk. Az alacsony illeszkedési statisztika rosszul működő kategóriát jelez, amelyben a modell elvárásaihoz képest kisebb a variabilitás.

3. Milyen mértékben érzékelhető a holdudvar hatás, vagyis egy értékelési szempont domináns szerepe?
Az előző kérdés eredményének további vizsgálata során az illeszkedési statisztika további problémákat is jelezhet: egyes kritériumok elhanyagolását illetve más kategóriák dominanciáját.

4. Milyen mértékben igazolják az eredmények az értékelőskála hat sávjának és négy szempontjának megfelelő működését?
A kategóriák illeszkedése, valamint a sávok valószínűségi görbéjének grafikus ábrázolása információt ad ezek alkalmazásáról, illetve feltárja az esetleges értékelői hibákat alkalmazásuk során.

4.2 Adatgyűjtés és –elemzés a második tanulmányban

A második tanulmányban az értékelő és az értékelőskála interakcióját vizsgáltuk különböző forrásokból származó adatok segítségével. Adott pontszámok, hangos gondolkodás módszerével és interjúval gyűjtött adatokat vetettünk össze a vizsgáztatói magatartás vizsgálata során abból a célból, hogy értékelői hibákat tárjunk fel. Ebben a tanulmányban tizenöt értékelő vett részt, tizenegy angol és négy némettanár. Az érzékelt értékelői magatartásra vonatkozó adatokat a releváns kategóriák szerint kódoltuk, majd következtetéseket vontunk le a többféle forrásból származó adat összevetésével. Az adatok elemzéséhez a tanulmánynak ebben a részében a MaxQDA2 (2005) programot használtuk. Ezekből az adatokból a következő kutatási kérdésekre kívántunk következtetéseket levonni:

1. Milyen tényezők váltanak ki a vizsgáztatókból olyan reakciókat, amelyek eltérítik őket az értékelőskála szerinti érékeléstől?
2. Milyen tényezők merülnek fel a javítás során, amelyek nem függnek össze a mérendő tulajdonsággal?

A 1. sz. táblázat összefoglalja a kutatási kérdések megválaszolása során alkalmazott adatgyűjtési és –elemzési módszereket.

1. sz. táblázat Módszertani mátrix

Kutatási kérdés	Adatok forrása	Adatelemzés módszerei
Mely értékelési szempontok váltanak ki elfogultságot az értékelőből? Melyik értékelési kritériumok esetében tapasztalható variancia hiánya? Milyen mértékben érzékelhető a holdudvar hatás, vagyis egy értékelési szempont domináns szerepe? Milyen mértékben igazolják az eredmények az értékelőskála hat sávjának és négy szempontjának megfelelő működését?	2011 írásműre analitikus értékelőskálán 27 értékelő által adott pontszám elemzése 7 időszakban, 3 éven keresztül gyűjtött adat	többitemparaméteres Rasch analízis FACETS programmal az értékelők, az érétkelési szempontok és a sávok illeszkedésének vizsgálatára
Milyen tényezők váltanak ki a vizsgáztatókból olyan reakciókat, amelyek eltérítik őket az értékelőskála szerinti érékeléstől? Milyen tényezők merülnek fel a javítás során, amelyek nem függnek össze a mérendő tulajdonsággal?	egyidejű hangosan gondolkodó módszer, valamint interjúk 15 értékelővel az értékelési folyamat és a feltételezett értékelői hibák vizsgálata céljából	kvalitatív adatelemzés MaxQDA programmal az értékelői hibák okainak feltárására

5 A főbb kutatási eredmények és jelentőségük

A kutatás céljaként tűzte ki, hogy empirikus bizonyítékokkal szolgáljon a BGF Nyelvvizsga és Továbbképző Központjában a középfokú írásfeladat értékeléséhez használt értékelőskála validitásának megállapításához. Célunk nem konstruktum validálás volt, hanem a mérőeszköz pszichometriai validitásának vizsgálata.

5. 1 Értékelőskála validitás

Az első kérdéskör a mérőskála működését helyezte középpontba.

1. Mely értékelési szempontok váltanak ki elfogultságot az értékelőből?

A FACETS analízis, melynek segítségével három év adatait vizsgáltuk két nyelvre vonatkozóan, jelzett elfogultsági értékeket, de szisztematikus elfogultsági mintát a vizsgált adatbázis nem mutatott. Ez az eredmény megerősíti azon korábbi tanulmányok eredményeit, amelyek hasonló módon nem tudtak rendszeresen visszatérő elfogultságot azonosítani az értékelők részéről. Annak ellenére, hogy a vizsgált adatokban szisztematikus hibát nem tudtunk azonosítani, mégis fontos ezt a vizsgálatot folyamatosan végezni, hiszen a nem szignifikáns, apró eltérések is fontos információval szolgálnak az értékelési folyamatról, és közvetett módon az egész mérési procedúráról. Az elfogultság lehet pozitív, vagy negatív attól függően, hogy az értékelő szigorúan vagy enyhén kezel-e egy-egy értékelési szempontot. Amikor az eredményeket értelmezzük, mindkét fajtáját az elfogultságnak szükséges vizsgálni függetlenül attól, hogy okoz-e hátrányt a vizsgázónak.

A második kutatási kérdés arra keresett választ, hogy az értékelők egyforma jelentőséget tulajdonítanak-e mindegyik értékelési szempontnak.

2. Melyik értékelési kritériumok esetében tapasztalható variancia hiánya?
Bár maga a kérdés is feltételez értékelői elhajlást, az eredmények nem erősítették meg a várakozásokat, mivel itt is a korábbiakhoz hasonló eredményeket kaptunk. Szisztematikus eltérést nem tapasztaltunk egyik kategória esetében sem. A kritériumok FACETS elemzése világos képet adott a kategóriák működéséről. Az illeszkedési statisztikák minden esetben az elvárt határértékek között mozogtak. Érdekes eredmény ugyanakkor, hogy az értékelők nem mutatnak következetességet túlzott szigorúság vagy enyheség területén az egyes értékelési szempontokra vonatkozóan. Ahogyan Eckes (2005) is megállapítja tanulmányában, az értékelők következetes szigorúak, illetve enyhék, de az egyes értékelési szempontokkal összefüggő szigorúságukban nincsen állandóság. Ez feltehetően arra utal, hogy a szempontok értelmezése szituáció- és feladatfüggő, és alátámasztja azt a feltevést, hogy minden értékelő, ugyan következetesen, de a maga sajátos módján értelmezi az értékelőskálát.

A harmadik és negyedik kérdés az értékelési kategóriákra és a sávokra vonatkozott.

3. Milyen mértékben érzékelhető a holdudvar hatás, vagyis egy értékelési szempont domináns szerepe?
Holdudvar hatás akkor tapasztalható, amikor az analitikus skála ellenére az értékelők inkább holisztikusan javítanak, mint globálisan. Az alacsony illeszkedési értékek holdudvar hatásra utalnak. Az összes vizsgált adat esetében az illeszkedési mutatók 0.66 és 1.53 között mozogtak. Csak ez a két szélső érték volt az előre meghatározott határértékeken kívül, több illeszkedési problémát a kategóriákra vonatkozóan nem találtunk. Az adatban szereplő egy alacsony kategória illeszkedési értéke egyetlen értékelő esetében nem jelent általános tendenciát a holdudvar hatásra. A elkülönítési mutató magas megbízhatósága is megerősítette, hogy az értékelők megfelelő módon el tudják különíteni a kategóriákat az értékelési folyamat során.

4. Milyen mértékben igazolják az eredmények az értékelőskála hat sávjának és négy szempontjának megfelelő működését?
A FACETS elemzés eredménye igazolta a hatfokú skála megbízható működését. A skála az az általánosan alkalmazott értékelőskála típus, amelyben a jobb teljesítményt magasabb pontszámmal, a gyengébb teljesítményt pedig alacsonyabb pontszámmal értékeljük. Ezért

tehát az alacsonyabb sávoknál alacsonyabb logit értékeket vártunk, a magasabb értéket képviselő sávoknál pedig magasabb logit értéket. Ez utóbbi állítás evidenciának tűnhet, de mindaddig feltételezés, amíg empirikus adatokkal alá nem támasztjuk. Eredményeink megerősítették a hatfokú értékelőskála megfelelő működését. Mind a szánadatok, mind pedig a grafikusan megjelenített adatok azt mutatják, hogy az értékelők jól elkülönítik a skála hat fokát nulla pont és öt pont között. A sávokkal asszociált logit értékek is megfeleltek az elvárásoknak, a monotonikusan növekvő skálaértékek fokai között az 1.4-es logit különbség egyetlen kivétellel mindenütt fennállt. Az elkülönítési mutatók is megerősítették a hat skálafok létezését.

5.2 Értékelői magatartás

Míg az első kérdéskör az értékelőskála megfelelő pszichometriai működését vizsgálta, a második kérdéskör a vizsgáztatói magatartást kívánta elemezni az értékelő és az értékelőskála interakciójának tükrében. Egyrészről javítás közben végzett egyidejű hangosan gondolkodás módszerével gyűjtött adatok, másrészről pedig az értékelőkkel készített interjúkból szerzett információk segítségével kívántuk az esetleges vizsgáztatói elhajlások okait felderíteni.

1. Milyen tényezők váltanak ki a vizsgáztatókból olyan reakciókat, amelyek eltérítik őket az értékelőskála szerinti érékeléstől?
Ugyan az első tanulmány nem bizonyította az értékelői elhajlás jelenlétét, ennek ellenére az interjúkból az derült ki, hogy egyes értékelési szempontok esetében a javítók elfogultságot éreznek. Ilyen a Feladatmegoldás kritérium, amelyet egyesek globális szempontként értelmeznek, és a másik három szempont összesítéseként kezelnek. Hasonlóan figyelmet érdemel a Nyelvhasználat szempontnál érzékelt holdudvar hatás, amely bizonyos értékelőknek a domináns szempont és esetlegesen negatív hatással lehet a többi szempontra is. Ezek a feltételezések, amint már említettük, nem nyertek számszerű megerősítést. Fontos azonban, hogy a felvetések szerint az értékelők tudatában vannak esetleges érzékeny értékelői pontjaiknak, amelyek értékelői hibához vezethetnek, és feltehetőleg mindent megtesznek annak érdekében, hogy ezek a negatív hatások ne tudjanak érvényesülni. Ezt a törekvésüket támasztja alá az is, hogy nagy többségük a kutatói kérés ellenére nem volt hajlandó a szempontokat fontosságuk szerint rangsorolni, mivel egyforma jelentőséget tulajdonítottak mindegyiknek. Ez utóbbi részben megcáfolja McNamara (1990) és Lumley (2005) állítását, mely szerint az írásfeladatok értékelése során a nyelvhasználat a domináns kritérium.

2. Milyen tényezők merülnek fel a javítás során, amelyek nem függnek össze a mérendő tulajdonsággal?
Az interjúk adatai alapján több olyan tényező merülhet fel a javítói magatartás vizsgálata során, amely érétkelői enyhéséghez vezet, mint amely szigorúságot idéz elő. Ugyan a normától semmilyen irányba nem kívánatos az eltérés, esetlegesen megjegyezhetjük, hogy azok az elhajlások, amelyek jutalmazzák a vizsgázót, feltehetően kevésbé károsak, mint azok, amelyek méltánytalanul bánnak velük. Pozitív értékelői elhajlást okozhat a vizsgázói intelligencia, amely érezhető a teljesítményből, a pozitív személyes tulajdonságok, valamint a kreativitás. Különösen érzékenyek az értékelők a másik oldalon a betanult szövegrészekre, a feladatból átemelt szövegrészekre és a kész formulákra.

5.3 Gyakorlati implikációk

Az eredmények két további kérdést vetettek fel, amelyek gyakorlati szempontból világítják meg a vizsgálat jelentőségét.

Milyen változásokat szükséges az értékelőskálán végrehajtani az eredmények tükrében?
Összességében a hatfokú, négy szempontú analitikus értékelőskála, amelyet középfokú írásfeladatok értékelésre használunk valid mérőeszköznek bizonyult. Ugyanakkor felül kell vizsgálnunk a Feladatmegoldás kritériumot, amely esetében változtatás látszik szükségesnek. A legtöbb vizsgáztatói elhajlás és elfogultság - még ha nem is szignifikáns mértékű – ezzel a kritériummal mutat összefüggést. A további adatok egyes sávleírások szövegezésében is változtatások szükségességét jelezték. Ezek a módosítások hozzájárulnak ahhoz, hogy a sávleírások egymástól függetlenül, önállóan is értelmezhetőek legyenek (Hawkey & Barker, 2004; North, 2000; Shaw, 2004). További finomítást igényel még a skála alsó fele, különösen a nulla és az egy pont közötti különbség markáns definiálása.

Az egész értékelési folyamatban milyen módosítások bevezetése szükséges ahhoz, hogy az értékelési szempontok interpretációjában nagyobb mértékű egyetértést tudjunk elérni?
Mivel az értékelőskálát az értékelési folyamat részeként kell vizsgálnunk, a validálási eljárásnak érinteni kell a mérési folyamat többi tényezőjét is. Az adatok fényében indokoltnak tűnik az egyéni javítói tulajdonságokra szabott javító tréning folyamat bevezetése. Egyértelműnek látszik az a feltételezés is, hogy az értékelői tulajdonságok nagymértékben függenek az aktuális feladattól, vagyis szituáció-függők. Ezért hát nem elégedhetünk meg a „képzett javító” fogalmával. Természetesen szükséges a javítókat ellátni azokkal az alapismeretekkel, amelyek általában szükségesek az írásfeladatok értékeléséhez. Nem mondhatunk le ugyanakkor a standardizációról, moderációról, vagy továbbképzésről, amelynek minden egyes javítói periódust meg kell előznie. Ezen az alkalmon kell kialakítani a közös, az adott feladatra vonatkozó értékelési keretrendszer, és benchmark minták segítségével kell meghatározni az adott sávban, szinten elfogadható vagy nem elfogadható jelenségeket. Fontos továbbá a vizsgáztatókról nyert információkat beépíteni az adminisztrációs folyamatokba is. Eltérő szigorúságú értékelőket célszerű értékelőpárként szerepeltetni, mert így elkerülhetjük azt, hogy két nagyon megbízhatóan, de egyformán szélsőségesen értékelő javító alkosson értékelőpárt. Az adatok információt nyújtanak az értékelői következetességről is. Az illeszkedés alapján következetlen értékelőt célszerű kihagyni a szubjektív értékelésű feladatok javításából. A több értékelő által is említett „rövidzárlat” jelenséget szintén megfelelő szervezéssel lehet kiküszöbölni.

6 Összefoglalás

A kutatás célja gyakorlatban alkalmazott mérőeszköz validálása volt azzal a szándékkal, hogy csökkentsük a szubjektív értékelésű feladatokkal asszociált pontatlanság feltételezését. Modern tesztelméleti eszközök lehetővé teszik, hogy azonosítsuk azokat a tényezőket, amelyek véletlen, vagy szisztematikus hibát okoznak mérési folyamatunkban. A teljesítmény értékelésének teoretikus modellje (McNamara, 1996), valamint Enghelhard (1992) íráskészség értékelési elmélete szolgált kutatásunk elméleti alapjául. Az adatelemzés nagyban támaszkodott a többitemparaméteres Rasch modell kínálta lehetőségekre, valamint Linacre (1989) vizsgáztatói elhajlás típusaira. Eredményeink empirikus adatok segítségével erősítik meg a mérőeszköz, valamint a mérési folyamat validitását. Várakozásainkkal

ellentétben nem sikerült értékelői elfogultságot feltárnunk, de ez csak további megerősítése a rendszer validitásának.

Vizsgálatunk két területen hozott fontos eredményeket. Ugyan szisztematikus hibákat nem tudtunk a rendszerrel kapcsolatosan azonosítani, kisebb, nem szisztematikus hibák előjöttek a vizsgálatok során, amelyeket szükséges javítanunk. A kutatás megerősítette annak szükségességét, hogy az alkalmazott módszert beépítsük tesztfejlesztői rendszerünkbe, és a folyamatos validálási procedúra részévé tegyük. Meggyőződhettünk másrészről a kvalitatív adatok hasznosságáról, amelyek betekintést nyújtanak a számadatok segítségével körvonalazott problémák mélyére.

7 A disszertáció felhasználhatósága

A kutatás során alkalmazott módszerek semmiképpen nem tekinthetők újnak a nyelvtudás mérése területén. Ugyanakkor az a forma, ahogyan a jelen kutatásban alkalmaztunk viszonylag ritkán alkalmazott, illetve dokumentált módszer. A szakirodalomban fellelhető többitemparaméteres vizsgálatok rendszerint keresztmetszetiek, nagy többségük egy adott időpontban vizsgálja az értékelőskála működését, leggyakrabban az értékelőskála kidolgozása után. Elengedhetetlen ugyanakkor, hogy az értékelőskála működését folyamatosan figyelemmel kísérjük, és az összegyűjtött adatokat rendszeresen elemezzük. Újdonságnak tekinthető továbbá az „értékelői validitás” fogalma, amely túlmutat a hagyományos értékelői megbízhatóságon. Az értékelői elhajlások szisztematikus vizsgálata ugyancsak jelentősen hozzájárulhat a mérési módszer hibáinak csökkentéséhez.

A vizsgálat tanulságai felvetik annak szükségességét, hogy a leírt validálási folyamatot a többi szubjektív értékelésű feladatnál is elvégezzük. Feltehetően egyes jelenségek általánosíthatóak a szóbeli vizsgák értékelési folyamataira is, és további vizsgálatokra van szükség, hogy ezeket a mérési folyamatokat is mélyreható pszichometriai vizsgálatnak vessük alá. Ugyan ezek a folyamatok idő-és forrásigényesek, mégis fontos elvégezni őket, hogy mérési folyamataink estében a lehető legnagyobb mértékű megbízhatóságot és érvényességet tudjuk biztosítani.

References
Hivatkozások

Carmines, E. G., & Zeller, R. A. (1979). *Validity and reliability.* London: Sage Publications Ltd.

Choppin, B. H. (1989). *Item banking and the monitoring of achievement*. Slough: National Foundation for Educational Research.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Hott, Rinehart, and Winston.

DeGruijter, D.N.M. (1984). Two simple models for rater effects. *Applied Psychological Measurement, 8(2),* 213–218.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A Many-Facet Rasch analysis. *Language Assessment Quarterly, 2(3)*, 197-221.

Engelhard, G. Jr. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education,* 5(3), 171–191.

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),* 93–112.

Hawkey, R., & Barker, F. (2004). Developing a common scale for the assessment of writing, *Assessing Writing,* 9(2), 122-159.

Kondo-Brown, K. (2002). An analysis of rater bias with FACETS in measuring Japanese L2 writing performance. *Language Testing*, *19*, 1-29.

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

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

Linacre, J. M. (2003-6). *A user's guide to FACETS: Rasch-model computer programs.* Chicago: MESA Press

Lumley, T. (2005). *Assessing second language writing*. Frankfurt am Main: Peter Lang.

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

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

North, B. (2000). *The development of a common framework scale of descriptors of language proficiency based on a theory of measurement*. New York: Peter Lang.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

O'Sullivan, B. & Rignall, M. (2001). *Assessing the value of multi-faceted Rasch bias analysis based feedback to raters for the IELTS Writing module.* Cambridge ESOL/The British Council/ IDA Australia: IELTS Research Report .

Saal, F. E., Downey, R. G., & Lahey, M.A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychometrical Bulletin, 88(2),* 413-428.

Shaw, S. D. (2004) IELTS writing: revising assessment criteria and scales. Phase 2. Retrieved August 6, 2004, from http://www.cambridge-efl.org/rs_notes/rs_nts15.pdf

Spearman, C. (1904). "*General intelligence", objectively determined and measured.* Retrieved September, 16, 2006 from http://psychclassics.yorku.ca/Spearman/chap1-4.htm

Viswanathan, M. (2005). Measurement error and research design. London: Sage Publications Ltd.

Wolfe, E. W., Moulder, B.C., & Myford, C.M. (2001). Detecting differential rater functioning over time using a Rasch Multi-Faceted rating scale model. *Journal of Applied Measurement, 2(3)*, 256-280