Bevezetés a méréselméletbe
(Introduction to Measurement Theory)

Linda Crocker
University of Florida

Fordította: Papp László
Lektorálta: Gál Attila

HÁTTÉR

A pszichológiának filozófiai gyökereiből tudománnyá fejlődése nagyrészt a kompetencia meghatározása és mérése együttes folyamatainak köszönhető. A pszichológiai kompetencia olyan elméleti változó, mely az oktatási folyamat eredményeként az egyén bizonyos módon való viselkedésének és teljesítésének valószínűségét jellemzi vagy azt, hogy hogyan viselkedik vagy teljesít egy *feladatdomainben*. (lásd Messick, 1989; Mislevy, 1996). A társadalomtudósok és nevelők gyakran tanulmányozzák a tudást, készségeket, képességeket vagy attitűdöket, melyek a kompetencia általános kategóriái. Konkrét példák a kompetenciákra: olvasásértés, tudományos attitűd, teljesítménymotiváció vagy éppen a zene iránti fogékonyság. A kompetenciákat operációs módon úgy határozhatjuk meg, hogy bemutatjuk mérésüket.

A mérés olyan folyamat, melynek során a tárgyakhoz, megfigyelésekhez, eseményekhez számértékeket rendelünk. E fejezetben a teszt szó azt a szisztematikus módszert jelöli, mellyel mintát veszünk egy egyén viselkedéséből egy bizonyos kompetenciaterületet képviselő strukturált feladatsorra adott válasza során. Ez a definíció magába foglal objektív teszteket, esszé jellegű teszteket, teljesítményértékelést, önértékelési naplót és megfigyelési mérlegeket. A teszteredmény értelmezése és ezen értelmezésen alapuló későbbi döntések szükségessé teszik, hogy a megfigyelt teszteredmény alapján következtetéseket vonjunk le a tesztfeladatok által képviselt megfigyelhetetlen kompetenciára vonatkozóan.

A méréselmélet rövid története

A teszteredmények alapján az egyének tudására, készségeire, képességeire következtetni ősi gyakorlat. Már i.e. 3000 évvel, az ókori kínai dinasztiák is alkalmaztak írásbeli vizsgákat a kormányzati posztokra pályázók kiválasztására. (Dubois, 1970). A régészek olyan agyagtáblákra bukkantak, melyek oktatók vizsgakérdéseit és azon tanítványaik válaszait tartalmazták, akik királyi írnokok szerettek volna lenni Mezopotámiában az i.e. az első ezredforduló idején. (Fishbein, 1981). Később a jezsuita papok a XII. században írásbeli esszéből álló vizsgákat tartottak a középkori európai egyetemeken, és a brit köztisztviselői rendszer is kidolgozott vizsgarendszert használt az egész birodalomban a XIX. században. (Dubois, 1970). E teszteredmények pontossága és hasznosíthatósága széles körben elfogadott tény volt, ami a vizsgáztatók képzettségébe vetett hiten alapult.

A hivatalos írásbeli vizsgák minőségének értékelésével kapcsolatos elméletek a XX. század elején kezdtek megjelenni, mint a fiatal pszichológiatudomány első szüleményei. A pszichológiatudomány számos úttörője (pl. Thurstone, Spearman, Pearson, Stoddard és Brown) fordult efelé a terület felé. 1904-ben William James, a neveléspszichológia megalapítója kezébe vette annak a tesztelméletről szóló első tankönyvnek ajándékkötetét, melyet volt doktorandusa, E.L. Thorndike írt. Thorndike azon való aggodalma ellenére, hogy mentora hogy fogadja majd ezt a grafikonokkal és képletekkel teli könyvet, James-nél lelkes

fogadtatásra lelt.(Joncich, 1968). A sikeres alkalmazás az első világháború során a hadsereg kiválasztási és osztályozó tesztjeinek kifejlesztésében bizonyságát adta a tesztelmélet hasznosíthatóságának. Ezt hamarosan széleskörű tesztelési programok követték a felsőfokú oktatásban a felvételik és tanulmányi teljesítmények mérése terén, (Allen & Yen, 1979), s ezáltal a fontos oktatási döntések kevésbé függtek már a családi kapcsolatoktól vagy az iskolák osztályozási rendszereiben rejlő különbségektől.

Természetesen a pszichológiai mérés mennyiségi megközelítése homlokegyenest ellentétben állt sok klinikai orvos nézetével, akik egy szubjektívebb, minőségi megközelítést részesítettek előnyben a diagnosztikus, szelekciós és csoportba sorolási döntéseik során. Meehl (1954) a következőképpen foglalta össze a helyzetet:

> Azok, akik gyűlölik az (objektív értékelési) módszert, mechanikusnak, összefüggéstelennek, *additívnak,* száraznak, mesterségesnek, nem reálisnak, önkényesnek hiányosnak…. rugalmatlannak, meddőnek, elméletinek, túlzóan leegyszerűsítőnek, áltudományosnak és vaknak tartják azt. Ugyanakkor, a klinikai módszert hívei dinamikusnak, átfogónak, jelentőségtelinek, *holisztikusnak*, árnyaltnak, ….. gazdagnak, mélynek, valódinak, érzékenynek, ….. természetesnek, valósághűnek és megértőnek nevezik (4.old.)

Mérföldkőnek számító metodológiai tanulmányában Meehl áttekintett több olyan több ezres mintával dolgozó tanulmányt, melyek összehasonlítást adnak az elbeszélgetésekre, megfigyelésekre és esettanulmányokra támaszkodó klinikai ítéleteken alapuló jóslások és a standardizált tesztadatokra alapozott jóslások pontosságáról. Az eredmények túlnyomórészt a teszteredmények használatát támasztották alá a klinikai ítéletekkel szemben olyan téren elért siker megjóslásában, mint a munkaköri képzés, értelmi betegségből való felgyógyulás és a büntető igazságszolgáltatás. Ennek eredményeként a tesztre támaszkodó méréselmélet a társadalomtudomány és oktatás szélesebb körében is tért nyert.

Gyakorlati követelmények

A méréselmélet gyakorlati alkalmazásának bővülésével a kapcsolódó irodalomban szakmai útmutatók jelentek meg a tesztfejlesztést, íratást, eredményértékelést és használatot illetően. A legismertebb e munkák közül a *Standards for Educational and Psychological Testing* [Az oktatási és pszichológiai tesztelés követelményei], mely már hatszor került kiadásra. (*American Educational Research Association, the American Psychological Association, & National Council on Measurement in Education,* 1999). A dokumentum célja, hogy „kritériumokat adjon meg a tesztek értékelését, a tesztelési gyakorlatokat és a teszthasználat hatásait illetően" (2. old.). 1954 óta, e könyvben hangsúlyos szerepet kaptak a tesztelmélet sarkalatos fogalmai, ami biztosítja, hogy az ajánlott gyakorlat elfogadott elméleten és empirikus kutatásokon alapszik.

A méréselmélet szükségessége: egy példa

A kompetenciaformálás, -mérés és tesztfejlesztés egyszerűsített illusztrációjaként képzeljük el a következőt. Egy diplomát szerzett hallgató, aki egy kilencedikes algebraórán végez megfigyelést, érdekes különbségeket kezd észrevenni a diákok között a tekintetben, hogy mennyire végzik a kiadott feladatokat (pl. olvassák a szöveget, dolgoznak a feladatokon, beütik számológépeikbe az adatokat, figyelnek a tanár utasításaira) vagy mással foglalkoznak (pl. barátaikkal beszélgetnek, videojátékokat játszanak, kinéznek az ablakon, vagy akár még

alszanak is). E megfigyelésekből a kutató megalkotja egy változó fogalmát, melyet "feladatorientáltságnak" nevez el.

Ezután kutatónk elhatározza, hogy e kompetencia tekintetében mérni fogja a diákokat. Lényegében, egy „feladatorientáltsági" tesztet akar összeállítani. E folyamat során több döntést is kell hoznia. Először el kell döntenie, milyen típusú adatokat akar gyűjteni és kitől. Például, megkérheti a tanárokat, hogy értékeljék a diákokat, kérheti a diákokat, hogy magukat értékeljék, vagy megfigyelheti és feljegyezheti, hogyan viselkednek a diákok. A megfigyeléses értékelést választja, ezért készít egy felsorolást különböző magatartásokról a megfigyelései alapján, melyek ezt a kompetenciát illusztrálják és tervet készít minden egyes diák összesen harminc perces, a félév során rövid időszakaszokban történő megfigyeléséről. Fel akarja jegyezni, mennyi időt tölt minden egyes diák a feladatnak megfelelő viselkedéssel az egyes megfigyelési időszakokban és össze akarja adni ezeket a bejegyzéseket, hogy megkapja a diák "feladatorientáltsági" pontját. E folyamat során a kutató megalkotta a kompetencia fogalmát, eldöntötte, hogy a mérési megközelítést fogja alkalmazni, amikor mintát vesz a diák viselkedéséből e kompetenciával kapcsolatosan, és még egy pontozási szabályt is alkotott. E ponton természetesnek tekinthetnénk, hogy a kutató alig várja, hogy itemeket hozzon létre, elkészítse a skálát és beavatkozzon az oktatásba az eredményeknek megfelelően. Ez azonban még korai lenne, előbb el kell készítenie és végrehajtania a mérések minőségének értékelési tervét is. Szerencsére jelentős anyag áll a rendelkezésére a méréselmélet és kutatás terén, hogy segítse értékelési törekvéseiben.

A MÉRÉSELMÉLET ELEMEI ÉS ALKALMAZÁSUK

A pszichológiai kompetenciák elvont jellege miatt, a teszteredmények csak közvetetten képviselik e kompetenciákat. Így minden tesztfejlesztő öt általános problémával találja magát szemben: (a) nem létezik egy általánosan elfogadott mérési megközelítés (b) a mérések csak egy korlátozott viselkedésmintát képviselnek (c) a mérésekben lehet hiba; (d) nincsenek pontosan meghatározott értékek a mérési skálán és (e) a kompetenciák valamilyen kimutatható logikai kapcsolatban kell, hogy álljanak más elfogadott kompetenciákkal vagy megfigyelendő jelenségekkel. A méréselmélet-tudomány célja, hogy megoldást találjon ezekre a problémákra és felbecsülje, hogy azok milyen mértékben befolyásolják a megfigyelt teszteredményeket egy felmérés során (Crocker & Algina, 1986). A méréselmélet-kutatást nagyrészt kategóriákba lehet sorolni a pontszerzés és tesztfelépítés, item analízis, reliabilitás és hibabecslés, teszteredmény-validitás modellje és az eredmények értelmezését segítő eszközök (pl. normafelállítás (*norming*), követelmény-meghatározás (*standard-setting*) és megfeleltetés/átszámítás (*equating*) alapján. E témák jelentése és alkalmazásuk a tesztek kifejlesztésében áll e fejezet középpontjában.

Domain-specifikáció, teszttervezés és itemelőállítás

A teszteredmények végső hasznossága nem a véletlenen múlik. Az alapot a tesztfejlesztés szolgáltatja. A tesztspecifikációk és itemek kialakításának folyamata általában a következőket foglalja magában:

1. A teszteredmények felhasználási céljainak meghatározása.
2. Azon kognitív képességek tartalmi és típuskategóriájának meghatározása, amelyre vonatkozóan a teszt felhasználói következtetéseket szándékoznak levonni (a kompetencia-domain meghatározása)
3. A különböző domainkategóriák súlyozása
4. Az itemek (feladatok) formátumának és a teszt hosszának meghatározása
5. A tesztíratás időhatárainak meghatározása

6. Itemírás és a forma kialakítása, utasítások írása (rubrikák készítése a teljesítményt mérő feladatokhoz)
7. Az itemek szakértő általi ellenőrzése (ne legyen maga az itemíró)
8. Az itemek szűk körű kísérleti tesztelése
9. Az itemek terepen való tesztelése széleskörű mintával és valódi tesztíratási körülmények között
10. Statisztikai itemanalízis a hibás itemek megtalálása, eltávolítása és kijavítása céljából.

E lépések bizonyos szintű kidolgozása segítséget nyújthat. Bár az itemírás elsősorban alkotásnak tekinthető, az itemíróknak mégis forrásanyagokra kell támaszkodniuk. A felmérendő kompetenciától függően, az itemek számára forrásanyagnak lehet tekinteni a közvetlen megfigyeléseket (ahogy a kutató a mi példánkban is tette), szakértő véleményeket, a kritikus incidensekről szóló jelentéseket, az alanyokkal, például vizsgázókkal készített interjúkat, munkatermékek, tantervek kritikáit, és irodalomkritikákat. Az ugyanazt az oktatási célt mérő több item kifejlesztésének szisztematikus megközelítései az 1970-es évek közepén jelentek meg először (Roid & Haladyna, 1982; Shoemaker, 1975), és számos tesztitem-írási útmutató létezik (pl. Haladyna, Downing, Rodriguez, 2002). És már felnőttkorba lépett annak a komputerszoftvernek a kifejlesztése, mely "itemklónokat" gyárt, melyek a felszínen különbözőek, de ugyanazt a tartalmat és folyamatot mérik (Irvine & Kyllonen, 2002). Az itemfejlesztés, ellenőrzés és kísérleti tesztelés alapos dokumentálására továbbra is szükség van a tesztfejlesztés folyamatának részeként (Downing & Haladyna. 1997).

Tereptesztelés és item-analizálási módszerek

Az itemek terepen való tesztelése lehetővé teszi a tesztfejlesztő számára, hogy megállapítsa, hogy az itemek megfelenek-e a kívánt nehézségi foknak és az elvárásoknak megfelelően működnek-e. Binet and Simon voltak az elsők, akik intelligenciatesztjük kidolgozása során itemeket választottak ki egy teszt összeállításához az itemek kísérleti elvégeztetése és a nyert adatok megvizsgálása céljára. (Baker, 1977). Az olyan itemek esetén, melyre 0 vagy 1 pontot lehet kapni, a klasszikus méréselméleten alapuló itemanalízis tipikusan a következő jellemzők vizsgálatából áll: itemnehézség, itemdiszkrimináció és az egyes érintett alcsoportok különböző itemteljesítése. A klasszikus tesztelméletben az itemnehézség az itemet helyesen megválaszoló vizsgázók arányában fejezhető ki és rendszerint *p*-vel jelölik. A normareferenciás tesztek esetén, közepes nehézségi szintű itemekre (pl. 0,30 – 0,70) van szükség az eredmények variabilitásának maximalizálására. A többfokú eredményskálás itemek esetén, mint pl. az esszéfeladatok vagy az értékelőskálás itemek esetén, az átlagos eredményt, a normától való eltérést, az egyes pontokban adott vizsgázói válaszok megoszlását vizsgálják. Ez azon a feltételezésen alapul, hogy fontosak a vizsgázók vagy vizsgázói csoportok közötti egyedi különbségek a mért kompetenciával kapcsolatosan és a tesztfejlesztő olyan itemeket szándékozik kiválasztani, melyek érzékenyen reagálnak ezekre a különbségekre.

Az itemdiszkrimináció arra utal, hogy az item mennyire tesz különbséget a magas eredményt és alacsony eredményt elérők között. Különféle statisztikai módszereket lehet alkalmazni az itemek diszkriminációs erejének értékelésére. Ezek közül a legegyszerűbb a *D,* diszkriminációs index. *A D*-t úgy számítják, hogy meghatározzák a teszt során nyilvánvaló különbséggel teljesítő alcsoportokat. Ezek az alcsoportok tipikusan a felső és alsó negyedet, felső és alsó harmadot, vagy a felső és alsó 27 főt jelentik a vizsgázók csoportjából. *(A D-t* ki lehet számítani az oktatott és nem oktatott csoportok esetén is.) Az egyes itemnehézségi fokokat (p-értékeket) ezekre a felső és alsó csoportokra számítják és a D-t kiszámolják minden egyes itemre:

$$D=p_U—p_L$$

ahol P_L az alsó szinten teljesítő csoport azon tagjainak aránya, akik jól válaszolták meg az itemet és P_U a felső szinten teljesítő csoport azon tagjainak aránya, akik jól válaszolták meg az itemet. Általában a 0,20-nál alacsonyabb értékű D esetén az itemet felülvizsgálják vagy törlik a tesztből. A negatív D-vel rendelkező itemek súlyosan hibásak és ki kell hagyni, vagy ellenőrizni kell, hogy nem a kulcs helytelen-e.

Ezen kívül diszkriminációs statisztikaként említhető a biszériális (kétsoros) pont és a biszériális korrelációs együttható, melyet egy kétértékű változó (0-val vagy 1-gyel pontozott item) folytonos változóval (pl. a teszt összpontszámával) való összevetésére fejlesztettek ki (lásd Allen & Yen, 1979; Crocker & Algina. 1986; Magnusson, 1967). Az esszé és teljesítményt mérő feladatok esetén, ahol a pontozás többfokú skálán zajlik, a *Pearson product moment* korrelációt számítják ki az item és a teszteredmény között, hogy értékeljék az item diszkriminációs erejét.

A vizsgázói alcsoportokkal szembeni igazságosságot lehet feltérképezni, főként kognitív képességek és teljesítmények esetén, a differenciált item teljesítménymutatók (DIF) vizsgálatával. Ezen elemzés célja tovább kell, hogy mutasson a különböző alcsoportok itemnehézségeinek összehasonlításán. Az itemigazságosság pszichometriai definíciója szerint a különböző alcsoportokban az azonos képességű diákoknak ugyanolyan valószínűséggel kell jól megválaszolni az itemet. Több analitikus technikát használnak a DIF megállapítására (Holland & Wainer, 1993; Langenfeld, 1997), de az egyik legegyszerűbb és leggyakrabban használt eljárás a Mantel-Haenszel eljárás (lásd Holland & Thayer, 1988).

A klasszikus itemstatisztika, mint például az itt leírt statisztika, mintaspecifikus. Ha az itemeket másodszor más képességmegoszlású mintán próbálják ki, az itemstatisztikai adatok megváltoznak. Tehát a terepen való tesztelés esetén a mintának megfelelően reprezentatívnak kell lennie arra a csoportra nézve, akik számára a teszt készült. Ugyancsak, ha itemeket küszöbölnek ki az itemelemzés során, akkor egy a korábbitól független mintán kell elvégezni a későbbi reliabilitás és validitás tanulmányokat, hogy ezáltal elkerüljük a validitás és reliabilitás-együtthatók túlbecslését.

A reliabilitás becslése

A leggyakrabban használt értelmében, a reliabilitás azt fejezi ki, hogy a mérések mennyire függnek össze vagy konzisztensek ugyanazon egyének esetén a különböző időpontokban vagy különböző fajta tesztek esetén. Például, amikor egy oktató szociológiai tanulmányokról szóló fejezet tartalma alapján összeállít egy tesztet, a teszt feltehető kérdések mintáját tartalmazza. Az oktató úgy gondolja, hogy egy másik, hasonló itemekből álló minta, mely ugyanazon a fejezeten alapszik, ugyanazt az eredményt nyújtaná ugyanazon vizsgázók esetén. Úgyszintén, amikor egy tanácsadó felmérést készít a diákok tanulmányi és esetleges szakmai érdeklődési területeiről, azt szeretné feltételezni, hogy a vizsgázók egy másik napon nem adtak volna jelentősen eltérő válaszokat. Szerencsére a tesztfejlesztőknek és felhasználóknak nemcsak egyszerűen elvégezniük kell a tesztet és reménykedni a legjobbakban az eredmény reliabilitását illetően. A klasszikus valós eredmény modell mennyiségi modell a konzisztencia (és hibavariancia) fokának becslésére egy adott teszteredmény-sorozaton belül. Traub (1997) felhívta rá a figyelmet, hogy a klasszikus tesztelmélet abból a jelentős hármas felismerésből született, hogy (a) a mérésben előfordulnak hibák, (b) a hibák véletlenszerűek és (c) két párhuzamos teszt eredményei közötti korrelációs együtthatót hibaindexként lehetne használni.

Háttérelmélet és a klasszikus valós eredmény modell képletei

A klasszikus valós eredmény modell és a reliabilitás fogalma Spearman kitartó érdeklődésének köszönhető, mellyel korrelációt akart vonni közvetlenül nem megfigyelhető jellemzők két hibát tartalmazó mérése között. A teszt tapasztalt eredményét (a nyers eredményt) két elméleti összetevő összegeként írta fel:

$X=T+e$,

ahol X a vizsgázó tesztben elért megtapasztalt eredménye, T a vizsgázó "valódi eredménye" és e véletlen hibakomponens, amely hozzájárul a tapasztalt eredményhez/pontszámhoz. Vegyük észre, hogy T nem tekinthető az egyén örök értékű tulajdonának, melyet mindenképpen megkapnánk függetlenül az egyén tesztelésének módjától. Ez az elméleti mennyiség az átlaga (várható értéke) minden lehetséges tapasztalt eredménynek, melyet ez a vizsgázó elérne, ha végtelenszer csináltatnák meg vele ugyanezt a tesztet (vagy ehhez a teszthez tökéletesen hasonló tesztet). A hibapont az eltérés a T és a vizsgázó tapasztalt X-e között bármely konkrét tesztelés során. Tehát egy adott tesztelés során sosem mondhatjuk meg bizonyossággal egyetlen személy valós eredményének értékét vagy az adott személy tapasztalt eredményében előforduló pontos hibamennyiséget.
Spearman felhasználta ezeket a meghatározásokat és feltételezések meghatározott körét, hogy megalkosson egy kifejezést egy vizsgázói csoport valódi és tapasztalt eredményeinek egy sorozata közötti korrelációra. Ez a korreláció r_{XT} a reliabilitás index. Spearman találmányának lényege, hogy ezt a mennyiséget két sorozat tapasztalt eredmény felhasználásával nyerjük, melyek ugyanazon vizsgázók két alkalommal két azonos vagy ugyanolyan teszt alapján történő tesztelésén alapulnak. E tapasztalt eredmények közötti korrelációt a *Pearson product moment* korrelációs formulával lehet kiszámítani. Ez a tapasztalt korreláció nyújt becslést egy elméleti mennyiségről, amit reliabilitás együtthatónak nevezünk. Spearman matematikailag bemutatta, hogy a reliabilitás együttható (r_{xx}) a reliabilitás index (vagyis a valódi eredmény és a tapasztalt eredmény korrelációjának) négyzete. Azt is bemutatta, hogy a reliabilitás együttható a tapasztalt eredmény varianciájának azon hányada, ami a vizsgázók valódi eredménye varianciájából adódik. Tehát, ahogy a tapasztalt reliabilitás becsült értéke az 1,00 felé közeledik, jobban megfelelnek egymásnak a vizsgázó valódi és tapasztalt eredményei.

Hibabecslés a tapasztalt teszteredményekkel kapcsolatban

Bár az egyes hibapontok nem megfigyelhetők, azonban lehetséges megbecsülni a pontatlanság fokát egy tipikus vizsgázó esetén a standard hibamérés alkalmazásával (SEM). A SEM a hibaeredmények disztribuciójának standard eltérésére ad becslést és úgy kaphatjuk meg, ha behelyettesítjük a tapasztalt eredmények standard eltérését és a reliabilitás együttható becsült értékeit a következő képletbe:

$$SEM = S_x\sqrt{(I - r_{xx})}$$

A SEM-et a vizsgázó tapasztalt eredménye körülötti megbízhatósági sávok becslésére használják: ez teszi lehetővé a teszt használója számára, hogy megállapítsa, hogy 68%-ban biztos, hogy a valódi eredmény ± 1 SEM-re van a vizsgázó tapasztalt eredményétől és hogy 95%-ban biztos, hogy a valódi eredmény ± 2 SEM –re van a vizsgázó tapasztalt eredményétől, ha feltételezzük, hogy minden vizsgázó esetén hasonló a hibadisztribúció.

A reliabilitásbecslés tanulmányok tervezése

Egy sorozat teszteredményhez nem csak egyetlen reliabilitásegyüttható tartozik (és nemcsak egy mérési hibastandard). Hanem az eredmények minden egyes sorozatához a reliabilitás becsült értékeinek egész családja tartozik, attól függően, hogy a tanulmány milyen tervet készít a reliabilitás becsült értékének kiszámolásához felhasznált adatok gyűjtéséhez. Öt főbb típusát különböztethetjük meg a reliabilitás tanulmányozásának:

1. A tesztelés-újratesztelés terv esetén a vizsgázók ugyanazon mintáján végeznek el egyetlen tesztet, melyet a szükséges idő elteltével újra ugyanazon a mintán megismételnek. Két sorozat eredmény közötti korrelációt nevezik tesztelési-újratesztelési együtthatónak. Ez az együttható a jellemző időbeli stabilitásáról ad becslést.
2. Az alternatív tesztlap terv a teszt kétféle tesztlap formájában való elkészítését követeli meg, melynek célja ugyanazon tartalmi területből való minta vétele és a kétféle tesztlap ugyanazon tesztspecifikációk alapján készült a tartalom, nehézség és itemformátum szempontjából. Mindkét tesztlapot ugyanazon a vizsgázói mintán próbálják ki és csak rövid szünetet tartanak a tesztek között. Ezután a két eredménysorozat között állítanak fel korrelációt, hogy megbecsüljék a két tesztlap ekvivalenciáját.
3. Az alternatív tesztlap, tesztelés-újratesztelés terv a két korábbi terv kombinációja.
4. A korrekciós felezési terv esetén a tesztet elfelezik (általában a páratlan számú itemeket az egyik, a páros számúakat a másik részhez osztják) a teszt együttes elvégeztetése után. Minden fél tesztet pontoznak és a két sorozat pontszám között korrelációt állítanak fel. A Spearman Brown korrekciót (melyről később lesz szó) használják a teljes hosszúságú teszt reliabilitásának becslésére.
5. Más egyszeri tesztírásra vonatkozó terveket is kidolgoztak, anélkül, hogy két részben pontoznák a tesztet, de a leggyakrabban használt közülük a Kuder Richardson 20 a 0-val vagy 1-gyel pontozott itemek esetén (Kuder & Richardson, 1937) vagy annak egy általánosabb formája, az alfa együttható (Cronbach, 1951), amelyet a többfokú skálán pontozott itemek esetén lehet használni (pl. esszé típusú itemeknél és az attitüdmérő itemeknél):

$$\alpha = K/(K-1)\left[1-\left(\sum S_i^2\right)/S_x^2\right]$$

Ahol K a teszt itemjeinek száma, $\sum S_i^2$ az egyes itemek varianciájának összege és S_i^2 a teszt összpontszámának varianciája. A belső konzisztencia együttható akkor használható, ha az itempontokat azon céllal adják össze, hogy megkapják a teszt összpontszámát és az itempontok tükrözik, hogy milyen mértékben adnak mintát az itemek a homogén tartalmi domainről. Az alfa együtthatót időnként az elméleti reliabilitás együttható alsó határaként értelmezik, de amint Brennan (2001a) és Traub (1997) megjegyzi, ez az értelmezés csak bizonyos megszorító feltételezésekkel együtt állja meg a helyét.

A reliabilitást befolyásoló tényezők

A felmérési helyzetekben számos tényező befolyásolhatja a tesztpontszám reliabilitásának becsült értékét. Először is, ha nincsenek különbségek a vizsgázó pontszámaiban, csökkenhet a reliabilitás becsült értéke. A különbségtartomány akkor szűkülhet le, ha a vizsgázói mintát egy összefüggő változó alapján előre megválogatják. Akkor is leszűkül a különbségtartomány, ha a teszt túl nehéz vagy túl könnyű a vizsgázók számára és így mindenki hasonló pontszámot kap. Másodrészt, a teszt hossza befolyásolja a reliabilitást,

mely rendszerint a teszt hosszával együtt nő és ellenkezőleg. A tesztek közötti hosszúságkülönbségekből adódó változásokat a Spearman Brown jóslási képlettel lehet megjósolni, $r_{xx}' = kr_{xx} / [1 + (k-1)r_{xx}]$ ahol *k* a teszt hosszának változásával kapcsolatos szorzó, r_{xx} pedig az eredeti reliabilitási együttható és r_{xx}' pedig a jósolt reliabilitás. Például, ha a teszt jelenlegi reliabilitásértéke 0,60 és a teszt hosszát megkétszerezzük, a kibővített teszt jósolt reliabilitása 0,75-re növekszik. Ha azonban háromszorosára növeljük a teszt hosszát, kisebb mértékű növekedését figyelhetjük meg a jósolt reliabilitás együtthatónak, 0,81 lesz. Ellenkezőleg, a teszt lerövidítésével általában csökken a teszt reliabilitása. Ez feszültséget okoz a tesztfejlesztéssel kapcsolatban, mert ahogy az itemek és válaszok összetettsége olyan mértékben megnövekszik, hogy esszé vagy teljesítménymérő formátumot kívánna, az egy bizonyos időintervallumban elvégezhető itemek számát csökkenteni kell. Ez alacsonyabb reliabilitáshoz vezethet, annak ellenére, hogy a felszínen a teljesítménymérés szorosabban megfelel az érintett viselkedési domainnek.

Továbbá, a vizsgázók tévesztése további véletlenségi varianciát eredményez a teszteredményekben, miáltal csökken a reliabilitás. A fáradtságból vagy a motiváció hiányából adódó figyelmetlenség szintén hibát eredményezhet a tesztpontszámban. Végül, ha a teszt írását sürgetik, vagyis a vizsgázók 10 vagy több százalékának nem jut ideje, hogy minden itemmel foglalkozzon, az megemeli a reliabilitásértéket, mert a lassan dolgozó vizsgázók rendszeresen pontot veszítenek az egy adott teszt során vagy tesztről tesztre el nem végzett itemeknél.

A reliabilitási együttható becslési alternatívái

Döntéskonzisztencia. A klasszikus valódi eredmény modellen alapuló reliabilitás-becsléseknek akkor van leginkább értelmük, amikor a cél az egyéni különbségek mérése az érintett kompetencia tekintetében. Ezt rendszerint normareferenciás mérésnek hívják. Sok oktatási felmérési helyzetben azonban az a cél, hogy meghatározzák egy vizsgázó teljesítményét egy előre meghatározott teljesítménykövetelményhez képest (pl. tehát az alapján osztályozni a vizsgázó teljesítményét, hogy felette vagy alatta van-e egy bizonyos pontszámnak az eredményskálán, ezt hívjuk kritériumreferenciás mérésnek. Ilyen esetekben a varianciaarányon alapuló klasszikus reliabilitási együttható kevésbé alkalmazható. Helyette a teszt felhasználójának fel kell mérnie milyen mértékben fognak a vizsgázók konzisztensen a részeredmények alapján csoportokba kerülni. Azon belátás, hogy a klasszikus tesztelmélet nem alkalmas a kritériumreferenciás tesztelés reliabilitásának megállapítására és az alternatív megoldások hátterében Hambleton, Swaminathan, Algina, és Coulson műve áll (1978). A kritériumreferenciás tesztek reliabilitásának számítási módszerével ma már számos mérési tanulmány (pl. Ebel & Frisbie, 1991) és gyakorlati feldolgozás Brennan (2001b) or Crocker and Algina (1986) foglalkozik.

Generalizálhatósági (Általánosíthatósági) elmélet. Amint a mérési feltételek összetetté válnak (mint pl. a többféle esszétémát többféle pontozó pontozza), nem lehet egyetlen becsült reliabilitásértékkel és az azt kísérő hibastandarddal meghatározni a szisztematikus variáció hatásait a mérési körülmények és a véletlenszerű hibavariáció tekintetében. Az általánosíthatósági elmélet vagy G-elmélet (*Generalizability theory*) (Cronbach, Gieser, Nanda, & Rajaratnam, 1972) kereteket nyújt a szisztematikus és hibavariancia összetevőinek definiálására és becslésére a komplex mérési formákban. Brennan (2001 a,b) és Shavelson és Webb (1991) ajánlhatók történeti forrásként az általánosíthatósági elmélet módszerének és alkalmazásának terén. (Lásd még Shavelson & Webb, 63. fejezet, ebben a kötetben).

Itemreakciós elmélet. Sok nagy tesztelési programban a klasszikus valós eredmény modellt a pontszám és itemelemzés terén felváltotta az itemreakcós elmélet (*item response theory* (IRT)) (Hambleton, 1989). Az IRT modellek egy matematikai függvényen alapulnak, melyet minden item esetében grafikusan lehet ábrázolni, és azt mutatja meg, hogy milyen valószínűséggel válaszolják meg helyesen az itemet a teszt által mért képességsáv bármely pontján. A tesztitem reakciósadatok IRT kalibrációja megadja (a) az tesztbeli itemteljesítmény hátterében álló statisztikai jellemzővel kapcsolatos becsült vizsgázói képességnek megfelelő pontszámokat és (b) ugyanezen a képességbecslési skálán az itemnehézség becsült értékeit (amely szám rendszerint -3.0 és +3.0 között van). Néhány modell ad becslést a diszkriminációs és tippelési paraméterekről is minden egyes itemre vonatkozóan. Az IRT modellek fontos eleme, hogy standard hibabecslést adjanak minden ponttal kapcsolatban a képességskálán. Egy másik fontos vonás az "itemparaméter változatlanság(invariancia)", amely lehetővé teszi a különböző formájú tesztek esetén (melyekben van néhány közös item) és különböző minták esetén a képességbecslés és itemnehézség egy skálán való kalibrálását, hogy összehasonlítsák azon vizsgázók teljesítményét, akik különböző formájú tesztet írtak vagy hogy egy bizonyos képességszinttel (követelménnyel) vessék össze egy vizsgázó teljesítményét az adott kompetenciaterületen. Az IRT módszerek számításai bonyolultabbak, több szigorú előfeltételezésen alapulnak és nagyobb méretű mintára van szükség, mint a hagyományos méréselméletnél. (Lásd 38. fejezet).

Validitás

A teszteredmények magukban semmit sem jelentenek. Inkább a teszt felhasználói által levont következtetések adnak nekik értelmet. A kritikus kérdés az, vajon lehet-e ezeket a következtetéseket igazolni. A validálás az a folyamat, melynek során meghatározzák a bizonyítás formáját és bizonyítékot gyűjtenek a következtetések igazolására (Cronbach, 1971). A XX. század első felében, a validáláshoz "szükség volt egy mérési kritériumra, amelyről azt feltételezték, hogy az megadja az adott változó 'valódi' értékét." (Kane, 2001. p. 319), és a validálás feladata az volt, hogy megmutassa, hogyan lehet e kritériumot megjósolni a teszteredményekből. Több mint negyven éve, Ebel (1961) a következőt jegyezte meg: "A validitás régóta az egyik legfőbb istenség a pszichometria panteonjában. Egyetemesen dicsérik, de kevés jó mű születik a nevében. A tesztvalidálás valójában sokak szerint a tesztfejlesztés legkevésbé kielégítő elemének tekinthető" (640. o.). Ebel úgy vélte, hogy a fő oka ennek a validitás többértelmű meghatározásában keresendő. E fogalom fejlődéstörténetének rövid áttekintése megerősíti ezt az állítást.

E fogalom definiálására tett első próbálkozások során, a mérésszakértők kijelentették, hogy a validitás arról szól, hogy "vajon egy teszt azt méri-e, amit mérnie kell" (American Psychological Association (Amerikai Pszichológiai Társaság, 1954)). Lassanként négy féle validitásról kezdtek beszélni, de később ezt a számot lecsökkentették három félére: (a) tartalmi validitás, (b) kritériumvaliditás és (c) kompetenciavaliditás:

1. *A* tartalmi validitásra úgy gyűjtöttek bizonyítékot, hogy dokumentálták a tesztfejlesztés folyamatát (pl. leírták azokat a lépéseket, amelyeket a tesztfejlesztők tettek annak érdekében, hogy definiálják a tartalmi domaint és biztosítsák, hogy az itemek ebből a domainből származnak) és domainbeli szakértőkkel ellenőriztették, hogy az itemek relevánsak-e és reprezentatívak-e a domaint tekintve.
2. Kritériumvaliditási bizonyítékot úgy gyűjtöttek, hogy a teszteredményeket olyan valós világbeli egy vagy két változóval kapcsolatos vizsgázói teljesítménnyel korreláltatták, melyekre a teszt felhasználói valószínűleg következtetni tudtak a teszteredményekből. Például, az egyetemi felvételi pontszámok alapján a teszt felhasználói a tanulmányi

sikerekkel kapcsolatosan akartak levonni következtetéseket. Így az egyetemi tanulmányi átlagot tekintik általában az ilyen tesztek kritériumváltozójának.

3. *A* kompetenciavaliditás bizonyítékát a teszt által mérni szándékozott jellemzőről alkotott pszichológiai elmélet sugallta. (Chronbach & Meehl, 1955). Legalább három féle empirikus bizonyítékot használtak a kompetenciavaliditás alátámasztására: (a) a faktoranalízis tudományából származó bizonyítékot, mely kimutatta, hogy a tesztitemek pontszámcsoportokat adnak meg, melyek az összetevők mögött rejlő kompetenciákról engednek elméletben fogalmat alkotni; (b) korrelációs tudományból származó bizonyítékot, mely azt bizonyítja, hogy a teszteredmények más jellemzők mérésének eredményével korreláltathatók, ahogy ezt a pszichológiai elmélet előfeltételezi. Campbell és Fiske (1959) hasznos keretet javasolt (a többjellemzős-többmódszeres mátrixot) annak megállapítására, hogy a teszteredmények vajon elsősorban a kiválasztott jellemzőt vizsgálják-e, vagy hogy jelentősen befolyásolja-e azokat az éppen alkalmazott mérési módszer.

1970-re, ahogy a validitás fogalmát úgy pontosították, hogy az a teszteredmények tulajdonsága (és nem a teszté) és a hangsúly a validálási folyamat felé tolódott (Cronbach, 1971). Az 1980-as években a validálást koherens érvelés felépítéseként határozták meg, mely logikán és empirikus bizonyítékokon alapul, melynek célja, hogy igazolja a teszteredményekből levont következtetéseket (Cronbach, 1988; Kane, 2001). Egyre inkább úgy tekintettek a validálás három féle megközelítésére, mint egy dolog egymással összefüggő összetevőire, melyeket a kompetenciavaliditás címszó alatt lehet összevonni. Cronbach (1988) a következőképpen foglalta össze a helyzetet: “A három különböző, de egyenértékű validitásról szóló, harminc évet megélt gondolat fölött már eljárt az idő” (4. o.).

Messick (1989, 1995) a kompetenciavalidálás hat egymást kiegészítő eleméről beszélt, melyek felválthatják a validitás “szentháromságát”:

1. A tartalmi reprezentáció szempontja (amelyet gyakran tartalmi validitásként emlegetnek) az item relevanciájának, reprezentatív jellegének és technikai minőségének megítélését alapul (Lennon, 1956; Sireci, 1998).
2. A szubsztantív szempont a feladatteljesítés hátterében meghúzódó bármely folyamatmodell elméleti magyarázatára összpontosít a vizsgázók által e folyamatokban felhasznált empirikus bizonyítékok mellett (Embretson & Gorin, 2001; Solano Fiores & Shavelson, 1997).
3. A strukturális szempont a felmérő feladatokkal és az azokat kísérő pontszámrubrikákkal vagy itempontszámsúlyokkal kapcsolatos belső szerkezetre követel a kompetenciadomain szerkezetének megfelelő, elméleti magyarázatot. (Benson, 1998; Loevinger, 1957).
4. Az általánosíthatósági (generalizálhatósági) szempont arra a bizonyítékra összpontosít, hogy az eredményértelmezés és a másféle pontszámtulajdonságok alkalmazhatók más populációk, helyzetek és feladattípusok esetén. A különböző alpopulációk esetén a differenciális kritériumjóslás tanulmányozása illusztrálja ezt a típusú validitásbizonyítékot csakúgy, mint a validitásgeneralizáció. (Hunter, Schmitt, & Jackson, 1982; Kane, 1982).
5. *A* külső szempont a teszteredmények és más kritériumok közötti viszonyok bizonyítékát foglalja magába, mind konvergens és diszkrimináns validitásbizonyítékokat, melyeket többjellemzős-többmódszeres tanulmányozással szerezhetünk meg. (Campbell & Fiske. 1959).
6. A konzekvenciás szempontja a kompetenciavaliditásnak azt teszi szükségessé, hogy a teszteredmény felhasználásával kapcsolatos döntések lehetséges és tényleges

következményeit mérjük fel. A szándékolt és nemkívánatos eredmények felmérése során figyelembe kell venni az igazságosság és a disztributív társadalmi igazságosság kérdéseit is. (Messick, 1980).

Bár a validálás első öt szempontja olyan módszereket tartalmaz, melyeket már régóta alkalmaznak a felmérő közösségben különféle címkék alatt, a hatodik, a konzekvenciális szempont ellentmondásosabbnak tekinthető. Amíg néhány kutató szerint fontos megvizsgálni a tesztfelhasználás hatásait a validálási folyamat szerves részeként (pl.: Shepard, 1997), mások szerint a tesztelés következményeinek vizsgálata túllép a teszteredmény validálásának körén és zavart okozhat (összezavarhatja a validálás tudományos céljait a szociálpolitikai kérdésekkel) (Mehrens, 1997; Popham, 1997; Tenopyr, 1996). Ez a vita a kompetenciavalidálás terjedelméről valószínűleg még a XXI. században is folytatódni fog, főként, mivel a számonkérési célú tesztelés egyre elterjedtebb. (Crocker. 2002).

Habár a validitáselmélet jelentős fejlődésen ment keresztül, azok, akik validálási tanulmányokat folytatnak vagy kritizálnak, főként a 3-6. szempont alapján, nagy valószínűséggel a teszteredmények és más változók közötti korrelációkkal dolgoznak. Ezért jó tisztában lenni azzal, hogy a validitás együtthatót negatívan befolyásolhatja (a) a jósló vagy a kritérium oldaláról a ponthatárok leszűkítése, (b) mérési hiba a jósló részéről vagy a kritériumpontokban, (c) nem megfelelő kritériummérés. Bár statisztikai kiigazításokat el lehet végezni, hogy felmérjük a határok leszűkítésének vagy az adatok megbízhatatlanságának hatásait, nincs korrekcióra lehetőség egy helytelenül vagy rosszul megválasztott kritériummérték esetén.

Normamegállapítás, követelményfelállítás és megfeleltetés (egyenlőségfelállítás)

Ha egy felmérés eredményét széleskörű tesztelési programokban használják fel csoportbeosztás, előléptetés, kiválasztás céljára vagy ezek alapján hoznak felelősségi döntéseket, lényeges szerepük van a tesztíratásról és pontszámértelmezésről szóló útmutatóknak. Fontos lehet még normatív eredménytáblázat elkészítése, vagy szükség lehet egy külön követelményt felállító tanulmányra, hogy meghatározzák, hogy az eredményskálán milyen minimum pontszámmal lehet teljesíteni a tesztet. Bármilyen követelményt felállító tanulmány elkészítésének folyamatát és eredményét dokumentálni kell. Végül, ha a különböző tesztformákat egymással felcserélhetőként kezelik a vizsgázókról hozott döntések során, akkor teszteredmény-megfeleltetési tanulmányokat kell végezni, hogy össze lehessen kapcsolni a különböző tesztformák nyers eredményeit.

A SZÉLESKÖRŰ TESZTELÉSI PROGRAMOKKAL ÉS JÖVŐBELI IRÁNYOKKAL KAPCSOLATOS KÉRDÉSEK

A múlt század első felében az informális osztálytermi felmérés és a standardizált teljesítménytesztelés békés egymás mellett élése volt megfigyelhető és a rendelkezésre álló információ jól szolgálta a tanárok, kutatók és kereskedelmi tesztfejlesztők igényeit. Később a felmérés jellege drámaian változni kezdett, főként annak a mozgalomnak az eredményeként, melynek célja az volt, hogy az iskolák és tanárok számot adjanak a diákok tanításáról. E mozgalom három fő mozgatója a következő volt:

1. 1965-ben a szövetségi alap- és középfokú oktatásról szóló törvény (Federal Elementary and Secondary Education Act (ESEA)) kötelezővé tette az I. címben szereplő programok formális értékelését. Ahogy a törvény végrehajtási szabályait

megalkották, a helyi iskolakerületeket is egyre inkább rákényszerítették, hogy olyan értékelési módokat fogadjanak el, amelyek erősen a standardizált teszteredmény-adatokon és normákon alapultak (Linn, 2000).

2. Az 1970-es és 1980-as években, egyre több állam tett kötelezővé minimális kompetenciatesztelést, hogy ez alapján léptessék át a diákokat magasabb osztályba és tegyék lehetővé számukra az érettségit. Ehhez a "kritériumreferenciás teszteléshez" (Nitko, 1980) előrehaladásra volt szükség a tesztelmélet terén, hogy útmutatást kapjanak a megfeleléshez szükséges pontszámok megállapításáról (Berk, 1986) és olyan tesztek kifejlesztésével kapcsolatosan, melyek közvetlenül kapcsolódnak a tartalmi tantervekhez
3. A 2001. évi "Egy gyermek sem maradhat le" törvény tekinthető a legelsöprőbb erejű szövetségi törvénynek az USA történetében a tesztelés terén, mely előírta, hogy 3-8. osztályban a tanulókat éves tesztelésnek kell alávetni, figyelemmel kell kísérni az előrehaladást és be kell számolni évente az egy év alatti előrehaladásról (AYP) és e törvény jelentős következményekkel járt az egyes diákok, tanárok és iskolák számára (lásd Linn, Baker, & Betebenner. 2002).

A közoktatás felett, a széleskörű tesztelési programok nagy jelentőséggel bírnak a főiskolákra, egyetemekre és szakmai képzésekre való bejutás szempontjából, és a bizonyítvány és engedélyszerzés szempontjából számos foglalkozás és szakma esetén, még akkor is, ha e folyamat hátterében rejlő tanok megkérdőjeleződnek (lásd Zwick, 2002). Végül, a kognitív pszichológia terén tett előrehaladás az elmélet újraformálásának szükségességéhez vezetett, a fókusz áthelyeződött a tesztbeli teljesítés alapján a mögöttes jellemzőről vagy teljesítményről levont következtetésekről a tudás szerkezetére és azokra a belső folyamatokra, melyek szükségesek azon feladatok végrehajtásához, melyek a tanítás eredményét jelentik (lásd Mislevy, 1996).

Ezen események elvezetnek a tesztelmélet és felmérés kérdéseivel kapcsolatos ismeretek bővítésének igényéig. Akik a méréselméletről többet szeretnének olvasni, a következő témákat és róluk szóló forrásokat találják:

1. Teszteredmény-átszámítás (lásd Kolen, 2004; Kolen & Brennan. 1995)
2. Teljesítménykövetelmények felállítása (pl. Cizek, 2001; Haertel, 2002)
3. Tesztbiztonság és szabálytalan reakciók (pl. Cizek, 1999; Wollack, 2003)
4. A tesztanyag tartalmi követelményeknek való megfelelésének értékelése (pl. Linn, Baker, & Betebenner, 2002; Bhola, Impara, & Buckendahl, 2003; Koretz & Hamilton, nyomdában)
5. Komplex itemek előállítása és itemklónozás (Irvine & Kyllonen, 2002)
6. Felvételi tesztelés, differenciált jóslás (Zwick, 2002) és validitásgeneralizációs tanulmányok (Hunter, Schmidt, & Jackson, 1982)
7. A csoportteljesítmény stabilitása és a változás mérése cohorscsoportok és longitudinális csoportok esetén (Brennan, Yin, & Kane, 2003; Linn & Haug, 2002; Yen, 1997)
8. A fogyatékos diákokra adaptált felmérés összehasonlíthatósága (Pitoniak & Royer, 2001) és a teszttartalom idegen nyelvre való fordítása (Sireci, 1997)
9. A validitáselmélet kiterjedése és tesztvalidálási módszerek (Crocker, 2003; Haladyna & Downing, 2004; Kane. 2001; Mislevy, 1996; Moss, 1998)
10. Komputeralapú tesztösszeállítás és íratás (Mills, Potenza, Fremer, & Ward. 2002; Parshall, Sprau, Kalohn, & Davey, 2003).

Ehelyütt csak ízelítőt adtunk olyan témákból és területekből, melyek a méréselmélet gyorsan változó tudománya keretein belül felderíthetők. A méréselmélet terén tett előrehaladás szorosan összekapcsolódik a modern oktatás tágabb oktatási, szociális, gazdasági és jogi kérdéseivel. Akik a tesztfejlesztés felmerülő technikai kérdéseit szeretnék kutatni vagy a nagy téttel járó felmérések politikai kérdéseit szeretnék megvitatni, azok számára elengedhetetlen a tesztelmélet alapos ismerete.

IRODALOM

Allen. N., & Yen. W. (1979). *Introduction to measurement theory.* Belmont. CA: Brooks-Cole.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington. DC: Author.

American Psychological Association. *(1954).* Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin. 51,* (2 Pt. 2). 1—38.

Anghoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 19—32). Hillsdale, NJ: Lawrence Eribaum Associates Inc.

Baker, F. (1977). Advances in item analysis. *Review of Educational Research, 47.* 151—178.

Benson, J. (1998). Developing a strong program of construct validation: A test anxiety example. *Educational Measurement: Issues and Practice,* 17(1**),** 10— 17, 22.

Berk, R. J. (1986). Performance standards on criterion referenced tests. *Review of Educational Research. 56,* 137—172.

Bhola, D. S., Impara, J. C.. & Buckendahl. C. W. (2003). Aligning tests with states' content standards. *Educational Measurement: Issues and Practice.* 22(3), 21—29.

Brennan, R. L. (2001a). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement. 38,* 295—317.

Brennan, R. L. (2001b). *Generalizability theory.* New York: Springer-Verlag.

Brennan, R. L. Yin, P, & Kane, M. T. (2003). Methodology for examining reliability of group mean differences. *Journal of Educational Measurement, 40,* 207—230.

Campbell, D. T., & Fiske. D. W. (1959). Convergent and discriminant validation by the multitrait—multimethod matrix. *Psychological Bulletin.* 56. 81—105.

Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it, and prevent it.* Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Cizek, G. J. (Ed.). (2001). *Setting performance standards: Concepts, methods, & perspectives.* Mahwah. NJ: Lawrence Erlbaum Associates Inc.

Crocker. L. (2002). Stakeholders in comprehensive validation of standards-based assessment. *Educational Measurement Issues and Practice, 21(1),* 5—6.

Crocker, L. (2003). Teaching for the test: Validity, fairness, and moral action. *Educational Measurement: Issues and Practice. 22(*3), 000—000.

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

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psvchometrika, 16,* 297—334.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. I. Braun (Eds.). *Test validity* (pp. 3—17). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Cronbach. L. J. (1971 **).** Test validation. In R. L. Thorndike (Ed.). *Educational measurement.* 2nd edition (pp. 443—507) Washington. D.C.: American Council on Education.

Cronbach, L. J. Gleser. G. C.. Nanda, H.. & Rajaratnam, N. (1972). *The dependability of behavioral measurements.* New York: John Wiley.

Cronbach, L. J.. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin. 52.* 281—302.

Downing, S. M.. & Haladyna, 1'. M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education. 10,* 61—82.

Dubois. P. (1970). *A history ofpsvchological testing.* Boston: Allyn & Bacon.

Ebel, R. L. (1961). Must all tests be valid? *American Psychologist. 16,* 640—647.

Ebel, R. L., & Frisbie, D. A. (1991**).** *Essentials of educational measurement* 5th ed.**),** Englewood Cliffs. NJ: Prentice-Hall.

Embretson, S., & Gorin. J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement. 38,* 343—368.

Fishbein. S. L. (1981). The Sumerians of Mesopotamia. In D. J. Crump (Ed.). *Splendors of the past: Lost cities of the ancient world* (pp. 34—71). Washington. D.C.: National Geographic Society.

Haertel. E. H. (2002). Standard setting as a participatory process: Implications for validation of standards-based accountability programs. *Educational Measurement: Issues and Practice, 21(1*), 16—22).

Haladyna. T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice. 23(*1), 17—27.

Haladyna. T. M., Downing, S. M., & Rodriguez. M. C. (2002). A review of multiple-choice item writing guidelines for classroom assessment. *Applied Measurement in Education, 15,* 309—334.

Hambleton. R. K. (1989). Principles and selected application of item response theory. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 147—200). New York: Macmillan.

Hambleton. R. K., Swarninathan. H., Algina. J., & Coulson, D. B. (1978). Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research, 48,* 1-47.

Holland, P. W. & Thayer, D. (1988). In H. Wainer & H. Braun (Eds.), *Test validity.* (pp. 129—141). Hillsdale, NJ: Lawrence Eribaum Associates Inc.

Holland. P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning.* Hillsdale. NJ: Lawrence Erlbaum Associates Inc.

Hunter, J. E.. Schmidt, F. L., & Jackson, G. B. (1982). *Meta analysis: Cumulating research findings across studies.* Beverly Hills, CA: Sage.

Irvine, S. H., & Kyllonen, P. C. (Eds.), (2002). *Item generation for test development.* Mahwah. NJ: Lawrence Erlbaum Associates Inc.

Joncich, G. M. (1968). *The same positivist: A biography of Edward L. Thorndike.* Middletown, CT: Wesleyan University Press.

Kane, M. T. (1982). A sampling model for validity. *Applied Psychological Measurement, 6.* 125—160.

Kane, M. T. (1992). An argument-based approach to validation. *Psychological Bulletin. 112,* 527—535.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement,* 38, 319—342.

Kolen, M. J. (2004). Population invariance in equating and linking: Concept and history. *Journal of Educational Measurement, 41*, 3—14.

Kolen. M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices.* New York: Springer Verlag.

Kuder. G. F., & Richardson, M. W. (1 937). The theory of the estimation of test reliability. *Psychometrika, 2.* 151—160.

Langenfeld, T. E. (1997). Test fairness: Internal and external investigations of gender bias in mathematics testing. *Educational Measurement: Issues and Practice, 16(*1), 20—26.

Lennon, R. T. (1956). Assumptions underlying the use of content validity. *Educational and Psychological Measurement. 16,* 294—304.
Linn, R. L. (1973), Fair test use in selection. *Review of Educational Research, 43,* 139—161.
Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29(2),* 4—16.
Linn, R. L., Baker, E. L, & Betebenner. D. W. (2002). Accountability systems: Implications of the requirements of the No Child Left Behind Act of 2001. *Educational Researcher, 31(6),* 3—16.
Linn, R. L., & Haug, C. (2002). Stability of school-building accountability scores and gains. *Educational Evaluation and Policy Analysis, 24.* 29—36.
Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports. 3.* 635—694 (Monograph Supplement 9).
Magnusson, D. (1967). *Test theory.* Boston: Addison-Wesley.
Meehl, P. E. (1954). *Clinical vs. statistical prediction.* Minneapolis: University of Minnesota Press.
Mehrens, W. A. (1997). The consequences of consequential validation. *Educational Measurement: Issues & Practice, 16(2),* 16—18.
Mehrens, W. A. (1998). Consequences of assessment: What is the evidence? *Education Policy Analysis Archives,* 6(13). 1—29.
Messick, S. (1 980). Test validity and the ethics of assessment. *American Psychologist, 35.* 1012—1027.
Messick, S. (1989). Validity. In R. L. Linn (Ed.). *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.
Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice. 14(4).* 5—8.
Mills, C. N., Potenza, M., Fremer, *3.* J., & Ward, W. C. (Eds.). (2002). *Computer-based testing.* Mahwah. NJ: Lawrence Erlbaum Associates Inc.
Mislevy, R. J; (1996). Test theory reconceived. *Journal of Educational Measurement, 34,* 379—416.
Moss, P, (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice, 17(2),* 6—12.
Nitko, A. J. (1980). Distinguishing the many varieties of criterion referenced tests. *Review of Educational Research, 50,* 461—485.
Parshall, C. G., Sprau, J. A., Kalohn, J. C., & Davey, T. (2003). *Practical considerations in computer-based testing.* New York: Springer-Verlag.
Pitoniak. M. J., & Royer, J. M. (2001). Testing accommodations for examinees with disabilities: A review of psychometric, legal, and social policy issues. *Review of Educational Research, 71,* 53—104.
Popham, W. 3. (1997). Consequential validity: Right concern—wrong concept. *Educational Measurement: Issues and Practice, 16(2),* 9—13.
Roid, G., & Haladyna, T. (1982). *A technology for test item writing.* New York: Academic Press.
Shavelson, R. J., &Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage.
Shepard. L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 10(2).* 5—8.
Shoemaker, D. M. (1975). Toward a framework for achievement testing. *Review of Educational Research. 45,* 127—148.
Sireci, S. (1997). Problems and issues in linking assessments across languages. *Educational Measurement: Issues and Practice, 16(1),* 12—19, 29.
Sireci, S. (1998). The construct of content validity. *Social Indicators Research, 45,* 83—1 17.

Solano-Flores, G., & Shavelson, R. J. (1997). Development of performance assessments in science: Conceptual, practical, and logistical issues. *Educational Measurement: Issues and Practice, 16(3),* 16-25.

Tenopyr, M. L. *(*1996, April). *Construct-consequences confusion.* Paper presented at the annual meeting of the Society of Industrial and Organizational Psychology, San Diego.

Traub, R. (1997). Classical test theory in- historic perspective. *Educational Measurement: Issues and Practice, 16(4),* 8—14.

Wollack. J. A. (2003). Comparison of answer-copying indices with real data. *Journal of Educational Measurement, 40,* 189—205.

Yen, W. (1997). Technical quality of performance assessments: Standard errors of percents of pupils reaching standards. *Educational Measurement: Issues and Practice. 16(3),* 5—15.

Zwick, R. (2002). *Fair game? The use of standardized admissions tests in higher education.* New York: Routledge Falmer.

Szószedet:

domain of tasks	feladatdomain	1. oldal
additive	additívnek	2. oldal
holistic	holisztikusnak	2. oldal
Standards for Educational and Psychological Testing	Az oktatási és pszichológiai tesztelés követelményei	2. oldal
item analysis	item analízis	3. oldal
reliability	reliabilitás	3. oldal
norming	normafelállítás	3. oldal
standard-setting	követelmény-meghatározás	3. oldal
equating	megfeleltetés/átszámítás	3. oldal
domain definition of the construct	kompetencia-domain meghatározása	3. oldal
item	item 3.,4. oldalon és több helyen; nem jelölt szöveg	
field testing	terepteszt elés (mezőtesztelés?)	4. oldal
item analysis	itemanalizálás	4. oldal
biserial	kétsoros	5. oldal
Pearson product moment	Pearson product moment	5. oldal
Generalizability theory	Generalizability theory	8. oldal
item response theory	item response theory	9. oldal
American Psychological Association	Amerikai Pszichológiai Társaság	9. oldal
Federal Elementary and Secondary Education Act (ESEA)	Federal Elementary and Secondary Education Act	12. oldal
beyond the K-12	a közoktatás felett	12. oldal
cohort groups	cohortcsoportok	12. oldal