Csapó Benő (1987): A kritérium-orientált értékelés. Magyar Pedagógia, 87. évf. 3. sz. 247-266.

CSAPÓ BENŐ

A KRITÉRIUM-ORIENTÁLT ÉRTÉKELÉS

A pedagógiai értékelés utóbbi másfél évtizedes történetének egyik legjelentősebb fejleménye a kritérium-orientált mérés elméletének megjelenése, az elmélet egyre kifinomultabbá válása és gyakorlati-technikai eljárásainak kialakulása.

Az elmélet magyarországi helyzete eléggé ellentmondásos. Legfontosabb mozzanatai viszonylag korán megjelentek egyes pedagógiai értékelőeszközök kidolgozásában. Más elemei, így elsősorban az adekvát matematikai apparátus

és általában az értékelési-adatfeldolgozási koncepció viszont még ma is csaknem teljesen ismeretlenek. A szakértők legszűkebb körén túl jellemző a félreértés és félreértelmezés. Olyan tesztekkel kapcsolatban is használják a kritérium-orientált jelzőt, amelyeknek kevés közük van a kritérium-orientált tesztelés elméletéhez vagy gyakorlatához.

A kritérium-orientált mérés nem csupán annyit jelent, hogy a tanulók tudását egy jól körülhatárolt területen mérjük fel, hanem ennél sokkal többet. A tesztelés új filozófiájáról van szó, a hagyományostól eltérő matematikai elméleteket alkalmazó adatfeldolgozási-elemzési eljárásokról, új tesztkészítési módszerekről. És ami mindezt szükségessé teszi: átalakulnak az oktatás módszerei, új oktatási stratégiák terjednek el, a tesztelés új funkciókat kap.

A technikai részletek kifejtése meghaladná e tanulmány kereteit, ezért a következőkben csak a kritérium-orientált értékelés legfontosabb kérdéseivel foglalkozom. Elsősorban a szemléletmódot szeretném bemutatni, a matematikai apparátusnak csak az egyszerűbb kérdéseit fogom érinteni. Az újszerű elemeket a klasszikus norma-orientált tesztelmélettel való kapcsolatokon, az azonosságokon és különbségeken keresztül mutatom be. Egy, a közeljövőben megírandó kézikönyvben részletesen szándékozom foglalkozni a kritérium-orientált tesztek készítésének elméleti hátterével, az itt alig érintett technikai kérdésekkel, a teljes adatfeldolgozó apparátussal.

A kritérium-orientált értékelés kialakulása és története

A "kritérium-orientált" terminust az angol szakirodalomban elterjedt "criterion-referenced" kifejezés fordításaként használom. Szó szerinti fordításban "kritériumvonatkozású", "kritériumra vonatkoztatott"-ként lehetne visszaadni. Néhány helyen az angolban is találkozhatunk a "criterion-oriented" formával, a németben pedig a "kriteriumsorientiert" (Fricke, 1974), más összefüggésben, de közel álló tartalommal: "lernzielorientiert" (Strittmatter, 1973) fordításban terjedt el. Találkozhatunk még a "domain-referenced", azaz (tudás-, illetve pszichikus) "területre vonatkoztatott" kifejezéssel is, melynek használatával kezdetben a "criterion-referenced" pontatlanságát kívánták kiküszöbölni. Ez a terminus nem terjedt el, ma már inkább csak akkor használják, ha a tudásterületre vonatkoztatottságot külön is hangsúlyozni akarják (Hively, 1974), egyébként a "criterion-referenced" szinonimája. Mindezeket figyelembe véve magyar pedagógiai terminusként a "kritérium-orientált" formát, rövidítéseként a "CR"-t, az eredeti angol rövidítést javaslom

A kritérium-orientált értékelés története a hatvanas évek elejére nyúlik vissza. A fogalmat először Robert Glaser használta 1962-ben, majd egy tanulmányában részletesebben is kifejti a kritérium-orientált és a norma-orientált (norm-referenced, NR) tesztelés közötti különbséget (Glaser, 1963). Elsősorban a pedagógiai értékeléssel foglalkozó kutatók vitái viszik tovább a gondolatot, több cikk, tanulmány (Popham and Husek, 1969; Ward, 1970; Popham, 1971a, 1971b; Hambleton and Novick, 1973) jelenik meg e kérdésekről. A tesztelés szemléletmódjának átalakulása, a formatív, diagnosztikai tesztek arányának növekedése, valamint a teljes elsajátítás igényével fellépő módszerek (mastery learning, personalized system of instruction) térhódítása nagymértékben inspirálta a megfelelő elméleti megalapozást. Csakhamar megjelentek az első monográfiák is. Ezek közül Popham (1978) könnyen követhető, tankönyvszerű formában (és amerikai szlengben) megírt, de a lényeges újítást jelentő matematikai-statisztikai kérdéseket csak alig érintő munkáját és a Berk (1980) szerkesztésében megjelent, lényegesen mélyebb kérdésekkel is foglalkozó könyvet emelném ki. A részletek kidolgozása a 70-es évek végétől rendkívül felgyorsult, a témakör irodalma az utóbbi években már szinte áttekinthetetlen burjánzásnak indult (a szakbibliográfiák a témakör közel ezer publikációját tartják számon).

A kritérium-orientált értékelés helyzete Magyarországon

Maga a kritérium-orientált értékelés szelleme az elméleti háttér kidolgozatlansága ellenére sem teljesen idegen a magyar kutatóktól. Szükségessége annyira nyilvánvalóvá vált, hogy bizonyos elemei Magyarországon is megjelentek, még mielőtt a nemzetközi mozgalom hullámai ideértek volna.

A pedagógiai mérések Magyarországon viszonylag rövid történetre tekinthetnek vissza, a nemzetközi mezőnyhöz képest jelentős fáziskéséssel indultak. A két világháború között egyfajta konzervatív beállítódás, majd a méréses módszerekkel (és általában az empirikus társadalomkutatással) szembeni ideológiai ellenállás akadályozta azt, hogy megfelelő szerephez jussanak. Ezzel is magyarázható, hogy mindmáig alig néhány szakember, illetve műhely foglalkozik elméleti igényességgel pedagógiai mérésekkel.

Az indulást minden bizonnyal (sok máshoz hasonlóan) Kiss Árpád nevéhez kapcsolhatjuk (pl.: Kiss, 1960—61, 1961), akinek munkája nyomán az OPI a pedagógiai mérések egyre fontosabb központjává vált.

A pedagógiai mérőeszközök kidolgozásában legjelentősebb lépésnek a Nagy József irányításával elvégzett munkálatokat tekinthetjük. A "Standardizált

Készségmérő Tesztek" című könyvsorozat és a "Standardizált Témazáró Tesztek" (STT) 18 kötete, amely a 70-es évek közepén az akkor érvényben levő felsőtagozatos tankönyvek mindegyikéhez elkészítette és reprezentatív mintán standardizálta a tananyag minden részletét magában foglaló témazáró teszteket, az egységes koncepció alapján készült pedagógiai mérőeszközök körében mind a mai napig egyedülállónak számít.

A kritérium-orientált szemléletmód legtudatosabban az STT elméleti hátterében (Nagy, 1972, 1975) és tesztkészítési gyakorlatában érvényesült, amely a tanulók tudását nemcsak egymáshoz vagy az átlaghoz, hanem a teljes tudásmennyiséghez viszonyítja. Figyelembe véve azonban az iskolai gyakorlat igényeit (és akkor még nem lévén kellően kidolgozva a kritérium-orientált mérés elméleti, matematikai-statisztikai háttere), ez a megoldás felemás volt: a tesztek a kritériumokhoz viszonyítva értékelték a tanulók tudását (% pontban), azonban az osztályzattá alakítás során a klasszikus, normatív tesztelés hagyományait követték.

Ez a megoldás az úgynevezett "standard osztályzat"-hoz (Nagy, 1973) vezetett, ami biztosította, hogy mindig a megfelelő, és minden tantárgy minden témakörében azonos arányban legyenek elégségesek, közepesek, jók, jelesek. Ez az arány a könnyű, országosan magas szinten elsajátított anyagrészeknél (az osztályzatok ponthatárai magasabbra kerültek) és a nehéz vagy gyengén elsajátított tananyagrészeknél (alacsonyabb ponthatárok) azonos volt. Vagyis a módszer egy kellően objektív matematikai eljárással tett eleget az iskola (illetve az egész oktatási rendszer és társadalmi környezete) azon elvárásainak is, hogy megfelelő arányban legyenek gyenge, közepes és jó minősítésű tanulók.

A standard osztályzat koncepciója természetesen meg is felelt az iskola tényleges működési mechanizmusainak, hiszen többnyire tényleg vannak, akik a tananyagot kevéssé, közepes vagy elfogadható színvonalon sajátítják el. Azonban ellentmondásossága azonnal kiderülne abban az ideális esetben, ha az iskola jól működne, és valóban meg is tanítaná mindenkinek azt, amit meg akar tanítani, hiszen akkor mindenkinek jelest kellene adni. (Az ideális határesetben a szórás hiánya miatt az egész statisztikai háttér összeomlana.) A tényleges gyakorlatban azonban az eredmények legfeljebb kissé ferde, ámde kielégítő közelítéssel normális eloszlást mutattak, ezért nem jelentett problémát a normatív tesztelés matematikai-statisztikai apparátusának használata sem.

A klasszikus (norma-orientált) tesztelmélet korlátai

A klasszikus tesztelmélet alapelveinek összefoglalásával egyben azt is áttekinthetjük, hogy milyen hiányosságok vezettek el a CR tesztek megjelenéséhez. Az NR tesztek alapfeltevése szerint a velük mérendő pszichikus tulajdonságok (akárcsak más, sok tényező által meghatározott tulajdonságok) a valószínűségelmélet centrális határeloszlás tételének megfelelően normális eloszlást mutatnak. Ha tehát ezekhez a tulajdonságokhoz hozzárendelünk egy mérőszámot, pontszámot (score), az is normális eloszlású lesz. Az adott tulajdonságot mérő tesztet úgy alkothatunk, hogy olyan ingeregyütteseket (az adott esetben tesztitemeket) állítunk össze, amelyekre adott válaszok a megfelelő mérendő tulajdonságtól függenek, azok által meghatározottak. Ily módon tesztfeladatként bármely ingeregyüttes felhasználható, ha az arra adott válasz a mérendő tulajdonság függvénye. Az már csak technikai kérdés, hogyan lehet az itemeket úgy súlyozni, könnyebb vagy nehezebb itemek válogatásával a tesztet úgy megalkotni, hogy az itemek együtteséből álló teszt eredményei már normális eloszlást mutassanak. Így például el lehet érni, csak a legismertebb példát említve, hogy különböző tesztek (mint például Raven kiegészítendő ábrasorai, vagy Wechsler többféle területről vett tesztfeladatai) felhasználásával megközelítően ugyanazt az intelligenciát mérjük.

Az így nyert mérőszámokat azonban valahogy interpretálnunk kell, hiszen önmagukban, minden viszonyítás nélkül nincs semmi értelmük. Ha tehát meg akarjuk mondani, hogy egy adott személy a vizsgált tulajdonság esetében elért x pontja mit jelent, azt valamihez viszonyítani kell. Elméletileg két alkalmas viszonyítási pontot is találunk. Megadhatjuk, hogy az adott egyén a vizsgált tulajdonság maximálisan lehetséges értékének (ha van ilyen) hány százalékával rendelkezik, vagy megadhatjuk azt, hogy más megvizsgált személyhez képest milyen pontszámot ért el.

A pszichológiai teszteket először olyan tulajdonságokra alkalmazták, amelyeknek nem határozhatjuk meg a maximális mértékét. Nem lehet tehát a mérések során nyert pontszámoknak úgy értelmet adni, hogy azokat a maximumhoz viszonyítsuk, például annak százalékában fejezzük ki. Ugyancsak nem rendelkezünk e tulajdonságok természetes nulla pontjával sem. Ezért nincs értelme azt mondani, hogy valaki az intelligencia maximumának x%-át birtokolja. Ebben az esetben nem tehetünk mást, mint a megvizsgált személyeket egymáshoz viszonyítjuk.

Az egymáshoz viszonyítás matematikáját a klasszikus tesztelmélet tökéletesen kidolgozta. Ennek a technikának a lényege az, hogy a méréseket egy

kellően nagy létszámú csoporttal (ezt referencia-csoportnak vagy normacsoportnak nevezzük) elvégezzük, majd a normális eloszlás elméletileg ismert tulajdonságainak, valamint a mérés során empirikusan kapott átlagnak és szórásnak a felhasználásával megadjuk, hogy az egyes egyedek a csoport átlagához viszonyítva hol helyezkednek el. Megadhatjuk akár százalékban kifejezve azt is, hogy az adott egyénnél a csoport hány százaléka nyújtott alacsonyabb teljesítményt, vagy egy alkalmas skálán mérőszámot rendelhetünk a vizsgált egyén teljesítményeihez. Azonban bármelyiket tesszük is, mindegyik ugyanazt fejezi ki: hol helyezkedik el az adott egyén a referencia-csoporthoz viszonyítva. A kapott adatok tehát relatív jellegűek, a normacsoporthoz viszonyítanak. (Innen ered az elnevezés: norm-referenced, azaz norma vonatkozású, normához viszonyított.)

A klasszikus tesztelmélet kimunkálta azokat a fogalmakat, és hozzá a megfelelő technikákat is, amelyekkel a tesztek jóságát jellemezhetjük, illetve ellenőrizhetjük. Matematikai-statisztikai hátterének legkiforrottabb összefoglalása Lord és Novick (1968) könyvében található. A három leggyakrabban használt fogalom az objektivitás, a reliabilitás és a validitás.

Az objektivitás az adatok felvételének, értékelésének és interpretálásának az adatfelvevő, -értékelő, -interpretáló személyétől való függetlenségének követelményét jelenti, ami empirikus úton viszonylag könnyen ellenőrizhető.

A reliabilitás fogalma már kissé bonyolultabb, a teszt megbízhatóságának mértékeként szokás meghatározni. Kissé szabadabban fogalmazva azt mondhatnánk, hogy a reliabilitás azt mutatja meg, a teszt mennyire jól méri azt, amit mér. Matematikai definíciója szerint egy teszt reliabilitása az adott tulajdonság valódi értéke (ezt pontosan nem ismerjük) és a teszttel mért értéke közötti korreláció négyzete.

Mivel a matematikai definícióban szereplő két változó közül az egyiket nem tudjuk mérni, az említett korrelációt, és így a reliabilitást sem tudjuk közvetlenül meghatározni. Lehet azonban valószínűségelméleti számításokkal olyan formulákat előállítani, amelyekről bizonyítható, hogy a reliabilitást alulról becslik (a valódi reliabilitás a becsült értéknél csak nagyobb lehet), és bennük csak mérhető, illetve a mért adatokból kiszámítható mennyiségek (többnyire az itemek és a teszt-összpontszámok átlaga és szórása) szerepelnek. Ezek a formulák mint a különböző reliabilitásmutatók ismeretesek. Közös jellemzőjük az, hogy mindegyik a tesztek belső konzisztenciáján alapul, vagyis azon, hogy a tesztek egyes itemjei mennyire ugyanazt a tulajdonságot mérik. Mindenekelőtt a Cronbach-féle alfa koefficienst, a különböző Kuder—Richardson-formulákat (leggyakrabban a 20-as és a 21-es számú használatos),

a Gulliksen-formulát stb. érdemes megemlítenünk. Ugyancsak a belső konzisztencián alapul a tesztfelezéses módszer, és igen hasonló a két teszt ekvivalenciájából kiinduló reliabilitásvizsgálat elvi megalapozása is. Bizonyos esetekben (ha a mérendő tulajdonság közben nem változik) lehet a reliabilitás számításának alapja a stabilitás is: ugyanazzal a teszttel két különböző időpontban elvégzett mérés eredménye közötti korrelációból következtethetünk a reliabilitás mértékére.

Az NR tesztek reliabilitásmutatói explicit vagy implicit módon mind feltételezik a teszteredmények normális eloszlását, és a reliabilitás akkor lesz magas, ha az egyes itemek megoldási aránya közel áll az 50%-hoz és szórásuk magas.

Az NR tesztek eleve akkor használhatók, ha magas szórást produkálnak, ha a vizsgált csoport tagjait jól "széthúzzák", ha jól differenciálnak. Ilyen szempontból azokat az itemeket, amelyek nem növelik a teszteredmények varianciáját, amelyeket majdnem mindenki meg tud oldani, vagy szinte senki, ki lehet hagyni a tesztből, csak a redundanciát növelnék.

A validitás, a teszt érvényessége azt fejezi ki, hogy a teszt azt méri-e, amit vele mérni szándékozunk. Ahhoz, hogy egy teszt validitása megfelelő legyen, egyben szükséges a magas reliabilitási érték is. Az azonban előfordulhat, hogy a teszt magas reliabilitása mellett is érvénytelen, vagyis valamit mér, mégpedig jól, de nem azt, amit vele mérni kívánunk.

A validitás elemzésére, a valid tesztek készítésére is sokféle technika áll rendelkezésünkre. Csak a fontosabbakat említve a prediktív, a tartalmi, a konstrukciós validitás biztosítása.

Az NR tesztek esetében a reliabilitásmutatók és a validálási eljárások is többnyire olyan értékeket vesznek alapul (közepes átlag, magas szórás), amelyek a pedagógiai tesztelés számára nem értékek, olyan feltételeket kötnek ki, amelyeket a pedagógiai tesztek használata során nem lehet biztosítani, illetve amelyeknek a betartása szándékainkkal ellentétes eredményekhez vezethetne. Így például már a kiinduló feltétel, az eredmények normális eloszlása sem érvényes, illetve az oktatás célja éppen az, hogy a célként megjelölt tudást a tanulók mindegyike magas szinten elsajátítsa.

Ennek megfelelően a normatív tesztek a pedagógia céljaira csak korlátozott mértékben, csak meghatározott funkciókra használhatók. E funkciók köre viszonylag pontosan körülhatárolható azzal, ha azt mondjuk, hogy normatív teszteket akkor kell használnunk, ha a tanulókat egymáshoz viszonyítva akarjuk értékelni. Például egy csoportból a legjobban (vagy leggyengébben) teljesítő, előre rögzített x%-ot akarjuk kiválasztani. Nem alkalmasak viszont a

pedagógiai értékelés legfontosabb funkcióinak betöltésére, a tanulás irányítására, a visszacsatolás biztosítására. A tanulás irányítását segítő formatív, diagnosztikai teszteknek ugyanis konkrétan kell kimutatniuk azt, hogy mi az, amit a tanuló már elsajátított, és mivel kell még foglalkoznia.

Az egyébként elegáns matematikai alapokon nyugvó klasszikus tesztelméletnek más természetű korlátai is vannak. E problémák megoldására kritérium-orientált értékelés mellett más irányú törekvések is vannak. (Néhányat ezek közül Horváth György (Horváth, 1985) is ismertet.)

A kritérium-orientált értékelés alapelvei

A pedagógiai értékelés egyik legfontosabb funkciója a tanulási folyamatok irányítása, és mint az előzőekből láttuk, a normatív tesztelmélet nem ad megfelelő alapokat az ilyen tesztek kidolgozásához. A pedagógiai folyamatok irányítására alkalmas teszteknek más tulajdonságokkal kell rendelkezniük. De melyek ezek a jellemző tulajdonságok? Ezeket összegyűjthetjük, ha megfontoljuk, hogy mi ezeknek a teszteknek a funkciója.

Kiindulhatunk abból a tényből, hogy a tesztek a gyakorlati használat során mindig valamilyen döntés megalapozására szolgálnak. A normatív tesztek ennek megfelelően alkalmasak arra, hogy a vizsgált egyéneket az adott tulajdonság szempontjából a csoportnormához (és ezáltal az egyéneket egymáshoz) viszonyítsák. Ezáltal segítenek eldönteni, hogy egy csoportból kik rendelkeznek az adott tulajdonsággal leginkább (vagy legkevésbé), kik tartoznak bele a legjobb vagy leggyengébb x százalékba.

Ezzel szemben a pedagógiai folyamatok döntő kérdése az, hogy egy egyén, függetlenül a társaitól, milyen mértékben rendelkezik a szóban forgó tulajdonsággal, és a tulajdonságnak ez a mértéke hogyan viszonyul egy minimális, maximális vagy más módon megadott optimális értékhez. A méréssel azt kell meghatároznunk, hogy az adott egyén adott tulajdonságának fejlődése/fejlesztése hol tart, mégpedig azért, hogy eldöntsük, melyek legyenek a további teendők. Kialakult-e egy tulajdonság a szükséges szinten, vagy további fejlesztésre van szükség; elsajátította-e a tanuló a tananyagot, vagy tovább kell azt tanulnia?

Ennek megfelelően a tesztek kifejlesztésének és jóságuk elemzésének is az lehet az alapja, hogy mennyire jól segítik ezt a döntést. A pedagógiai tesztelés ma még oktatáscentrikus, a tesztek többsége az oktatás eredményesebb irányítását segíti. A tágabb értelemben vett személyiségfejlesztés szolgálatába állítható affektív tesztekre azonban ugyanúgy alkalmazható a CR megközelítés.

A kritérium-orientált tesztek kifejlesztésének fázisai

A CR tesztek kidolgozásának kialakultak bizonyos eljárásai. Ezek az eljárások lényegesen különböznek a normatív tesztek kidolgozásának technikájától.

A kiinduló lépés általában egy deskriptív séma kidolgozása. Ennek a sémának kell összefoglalnia a mérendő tartalom minden lényeges elemét. Elkészítésére nincs általánosan használható technológia, de léteznek az esetek többségében jól használható eljárások.

A tudásszintmérő tesztek esetében a legprecízebb eljárás a Nagy József (1972) által kidolgozott megoldás, amely a tananyagban levő tudáselemek számbavételén alapszik. A deskriptív séma ez esetben a tanyanyagban szereplő fogalmak és tények teljes rendszere. A totalitás elve (tehát az, hogy minden egyes tudáselem bekerül a tesztbe) kizárja az önkényes válogatásban megnyilvánuló szubjektivitást. Egy rögzített tananyag fogalmainak és tényeinek a tesztbe való leképezésére ez a ma ismert legjobb megoldás. Problematikusabb a kognitív műveleteknek nevezett, készség jellegű tudáselemeknek a tananyagból való levezetése. Erre is ad azonban egy meglehetősen objektív megoldást az alsó tagozatos szöveges feladatbank (Csáki—Nagy, 1976).

Ha azonban nem az érvényben levő tankönyvet tekintjük a tesztelés viszonyítási alapjának, akkor a deskriptív séma felállítására csak általános alapelveket lehet megfogalmazni. Ezek közül is a legfontosabb a teljes struktúra feltárásának elve. Ebből már lényegében következik a másik alapelv is, mégpedig az, hogy CR teszteket többnyire csak egy szűkebb terület vizsgálatára lehet kidolgozni. Csak így lehet a szelektálásból fakadó kényszerű szubjektivitást elkerülni. A teljes struktúra felállítására ritkán lehet egyértelmű megoldást találni. Többnyire csak akkor, ha a mérendő tulajdonság néhány egyszerű változó néhány értékével jellemezhető, és a teljes struktúrát ezek kombinatorikailag képezhető változatai alkotják. Más esetekben a deskriptív séma a szaktudományoknak vagy a pszichológia eredményeinek elemzése révén állítható fel, és végső soron a szakértők konvenciója szentesíti. Ez azonban már a méréselméletből kivezető, részben a tantervelmélet és a célelmélet körébe utalható kérdés.

A deskriptív séma szolgálhat alapul a tesztitemek elkészítéséhez. A sémának olyannyira egyértelműnek kell lennie, hogy annak alapján bármely kompetens szakértő egyértelműen el tudja készíteni a tesztfeladatokat.

A klasszikus tesztelmélet kulcsfogalma a homogenitás, ami azt jelenti, hogy minden item megközelítőleg ugyanazt a tulajdonságot méri. A CR tesztektől ezt nem lehet elvárni, az egyes itemei, habár ugyanannak a területnek a

részeit vizsgálják is, ezek a részek lényegesen különbözhetnek is egymástól. Így a homogenitás nem lehet egy CR teszt értékmérője. A klasszikus homogenitás helyébe egy másik homogenitás-fogalom lép, a levezetés homogenitása (derivatív homogenitás). Ezen azt értjük, hogy a deskriptív sémából minden esetben azonos módon, a nehézséget és a bonyolultságot visszatükrözve vezessük le a tesztitemeket. Ha a kétjegyű számok szorzását a "87 x 46 = ?", a háromjegyűekét pedig a "100 x 100 = ?" feladattal akarjuk tesztelni, akkor egészen biztosan nem tettünk eleget a derivatív homogenitás követelményeinek.

A tesztfeladatírás helyességének ellenőrzésére a klasszikus homogenitást is fel lehet használni, és a feladatok helyességét empirikus módszerekkel is lehet ellenőrizni. Ha ugyanis ugyanannak a deskriptív sémának a felhasználásával több szakértővel készíttetünk tesztfeladatokat, az egymásnak megfelelő feladatoknak most már a klasszikus értelemben is homogén tesztet kell alkotniuk, és azonos nehézségűnek kell lenniük.

A kritérium-orientált itemanalízis

A CR tesztek jóságának megítéléséhez is a klasszikustól eltérő módszereket kell választanunk. Az itemanalízisnek két alapvető megközelítésmódja lehetséges. Elemezhetjük az itemeket még az előtt, hogy azokat bármilyen mintán kipróbáltuk volna, és elemezhetjük egy elegendően nagy minta által produkált megoldások adatainak statisztikai elemzésén keresztül. Az egyszerűség kedvéért nevezzük az egyiket a priori, a másikat a posteriori megközelítésnek. Az NR tesztek elemzése során az a posteriori megoldások domináltak, elsősorban azért, mert a kidolgozott matematikai eljárások ezt lehetővé tették, másrészt pedig, mert a vizsgált tulajdonságok és a tesztitemek közötti kapcsolat bizonytalansága az a priori elemzéseket eleve kizárta. A CR tesztek esetében azonban a mérendő tulajdonság többé-kevésbé jól definiált, és ez módot ad az itemek megfelelőségének elemzésére.

A CR tesztek jóságmutatói többnyire már nem is a tesztre vonatkoznak, hanem arra, hogy a teszt alapján meghozott döntések mennyire jók. Így beszélhetünk a döntés megbízhatóságáról vagy validitásáról.

Validitás

A CR tesztek validitásának vizsgálatában két fő tendencia figyelhető meg: 1. a vizsgált területnek való megfelelés és 2. a teszt alapján hozott döntés érvényessége. A vizsgált területnek való megfelelést többnyire a pri-

ori, a döntés érvényességét statisztikai (a posteriori) módszerekkel vizsgálhatjuk.

A CR tesztek validitásának mérlegelése látszólag egyszerű feladat, hiszen van egy meghatározott pszichikus terület, tulajdonság, amit mérni akarunk, és azt kell csupán megállapítanunk, hogy a teszt valóban megfelel-e az adott területnek. Ennek meghatározásában azonban a statisztikai módszerek általában háttérbe szorulnak. Az itemek és a mérendő tartalom közötti megfelelés (item-objective congruence) biztosítása a megfelelő tesztkészítési technika kérdése, és ez esetben a validitás vizsgálata az a priori elemzések körébe utalható. A validitás elemzésének ez a megközelítése különösen érvényes azoknál a teszteknél, amelyeknél a "criterion-referenced" név inkább a területre vonatkoztatottságot (domain-referenced) jelenti. A validitás a deskriptív séma helyességén és az itemek levezetésének pontosságán múlik. Mindkét esetben kulcsszerepe van a szakértői elemzéseknek. Természetesen az itemek megfelelőségéről a szakértőktől gyűjtött adatoknak az egzakt analízisére is lehet eljárásokat, statisztikai mutatókat kidolgozni (Hambleton, 1980).

A döntési validitás vizsgálatára már alkalmazhatóak az a posteriori statisztikai vizsgálatok is. Itt azonban mindig konkrétan meg kell állapodni abban, hogy milyen döntés érvényességének vizsgálatáról van szó. Mivel a legtöbb tesztet oktatási kontextusban használják, a döntés két tényezőre irányulhat. Dönthetünk magáról az oktatásról, megvizsgálva annak hatékonyságát, szelektálhatunk a hatékony és a kevésbé hatékony megoldások, módszerek között, meghatározhatjuk egy-egy oktatási folyamat gyenge, megerősítésre szoruló pontjait. Gyakrabban van azonban szükség annak eldöntésére, hogy ki mit sajátított már el, és kinek mit kell még tanulnia.

Mindkét esetben akkor lesz valid a döntés, ha valóban azt tudjuk meghatározni, hogy az oktatás adott szakasza mennyit tett hozzá a tanulók tudásához, és nem kívánjuk értékelni a már korábban más forrásból megszerzett tudást. Ekkor a tesztnek az oktatással szemben kell érzékenynek lennie (instructional sensitivity), és azok az itemek növelik a teszt validitását, amelyek olyan elemeket vizsgálnak, amelyeket az oktatás előtt a tanulók többsége nem tudott.

Számos indexet fejlesztettek ki, amellyel az itemek oktatással szembeni érzékenységét jellemezhetjük, mindegyik a klasszikus diszkrimináló erő indexével mutat rokonságot. Itt csupán három könnyen kiszámítható és interpretálható indexet mutatok be. A következőkben p az itemet jól megoldó tanulók arányát jelöli, tehát $p = j/n$, ahol a j a jó válaszok száma, n pedig az ösz-

szes válasz száma (a minta elemszáma). Így p egyben az item nehézségi indexe is.

Az oktatás előtti (oe) és az oktatás utáni (ou) különbségen alapuló index:

$$D(ou-oe) = p(2)-p(1),$$

ahol p(2) az oktatás utáni, p(1) az oktatás előtti mérésre vonatkozik. Látható, hogy annál magasabb egy item D(ou—oe) indexe, a tanulás utáni és a tanulás előtti nehézségi indexek különbsége, minél inkább az adott tanulási folyamat eredményeként létrejött tudást méri az item.

Az adott tudás szempontjából oktatott és nem oktatott (két hasonló) csoport közötti különbségre hasonló indexet alapozhatunk, és így kikerülhetjük a második tesztelésnél a teszt ismerősségéből fakadó nemkívánatos effektusokat:

$$D(o-no) = p(o)-p(no),$$

ahol p(o) az oktatott, p(no) a nem oktatott csoportra vonatkozik.

Az oktatás előtti és az oktatás utáni eredményekből indul ki az egyéni tudásnyereséget kifejező index:

$$D(et) = p(-, +),$$

ahol p(-, +) azoknak az aránya, akik az oktatás előtt rosszul, az oktatás után pedig jól oldották meg az itemet.

Érdemes megfigyelni, hogy a kritérium-orientált szemlélet technikai háttere is mennyire a változások vizsgálatára épül, és mennyivel alkalmasabb pedagógiai jelenségek tanulmányozására, mint a pszichikum statikus felfogására épülő, a stabilitással számoló klasszikus tesztelmélet.

Az első két index negatív is lehet, de csak akkor, ha az oktatás hatására kevesebb tanuló oldja meg jól az itemet. Egyébként minél nagyobb az index (minél közelebb áll a +1-hez), annál inkább olyan tudáselemet mér az item, amit az adott oktatás eredményez. Sok hasonló indexet ismerünk a CR tesztek itemjeinek vizsgálatára (Berk, 1980, 60—63.), ezeknek a kiszámítása azonban már többnyire nagyobb statisztikai mintát tételez fel, és számítógép használatára van szükség.

Néhány, az NR tesztek elemzésére kidolgozott módszer és számítógépes eljárás kis technikai módosítással, de egészen más funkcióval alkalmazható a CR tesztek vizsgálatára is. Egyik ilyen módszer a pontszámkülönbségen alapuló, és a klasszikus tesztanalízis item-tesztösszpontszám korrelációjának mintájára kiszámított korrelációs együttható.

Minden tanulónál mind az itemekhez, mind pedig a teljes teszthez hozzárendelünk egy változás-pontszámot (change-score): az oktatás utáni tesztelés és az oktatás előtti tesztelés során elért pontszámok különbségét. Itemek esetében (di) ez -1, 0, vagy +1 lesz, a teszteknél (dt) a két pontszám különbségeként előálló, többnyire pozitív szám. Majd ezekkel a különbségpontokkal kiszámítjuk az r(di, dt) item-tesztkorrelációkat. Mivel a teszt különbségpont azoknál a tanulóknál lesz magas, akik az oktatás hatására sokat tanultak, a magas r(di, dt) azokat az itemeket jelöli ki, amelyek legjobban kifejezik az oktatás hatását.

Ezek a statisztikai mutatók kiküszöbölik az NR itemanalízis korlátait, azt, hogy csak olyan változókra működik, amelyeknek elég magas varianciájuk van. Az első három index azáltal, hogy közvetlenül a gyakoriságokkal számol, az utóbbi pedig úgy, hogy ha az oktatás utáni eredményeknek már nincs is számottevő szórásuk, de az oktatás előtti eredmények szóródtak, akkor a különbségpontoknak is lesz szórásuk. A kétféle megoldás alkalmazhatóságának feltételei is különböznek: a bemutatott három index akkor éri el a maximumot, ha az oktatás előtt senki nem tudja az adott itemet megoldani (nincs variancia), a különbségpont-korreláció viszont feltételezi az oktatás előtti eredmények szóródását.

Az oktatás, a tanulás irányítása során többnyire azt kell eldöntenünk, hogy a tanuló elsajátított-e egy tudáselemet, és továbbhaladhat a következő tanulási egységre, vagy pedig nem sajátította még el, ezért azt tovább kell tanulnia. A teszt vagy tesztitem eredménye alapján meghozott döntés akkor érvényes, ha csak azokat a tanulókat (de azokat mindet) engedjük továbbhaladni, akik az adott elemet valóban elsajátították. A teszt, tesztitem tehát akkor valid, ha annak megoldásához éppen az adott tudáselemek szükségesek, és nem elegendő valamilyen általános értelmesség. Mint láttuk, a fenti eljárásokkal éppen azokat az itemeket tudjuk kiválogatni, amelyek eleget tesznek ezeknek a feltételeknek, sőt, az oktatással szemben kevéssé érzékeny feladatokat kihagyva azoktól a feladatoktól szabadulunk meg, amelyeket az adott oktatás nélkül is vagy annak hatására sem tudnak megoldani a tanulók.

Ha a normatív tesztelmélet alapján javítjuk a teszteket, és sorra kihagyjuk a teszt összpontszámával (vagy egymással) csak kismértékben korreláló elemeket annak érdekében, hogy növeljük a reliabilitást, fennáll az a veszély, hogy a nagyon is konkrét és éppen a megtanítani kívánt tudáselemeket vizsgáló itemek hullanak ki. Ezért végül nem marad más, mint az általános értelmességet legjobban reprezentáló, de esetleg a vizsgálni kívánt szférával legkevésbé összefüggő homogén massza. A reliabilitás növekedésével tehát

egyben a validitás csökken, valamit egyre jobban mérünk, de ez a valami kevésbé az, amit mérni akartunk. A CR itemanalízis validálási eljárásai éppen fordított eredménnyel járnak: alkalmasak arra, hogy segítségükkel az adott oktatási folyamat során megszerezhető tudást legjobban reprezentáló itemekből állítsuk össze a tesztet.

Reliabilitás

A CR tesztek reliabilitásának vizsgálatában két fő megközelítés figyelhető meg. Az egyik a CR tesztek segítségével végzett döntések megbízhatóságát veszi alapul. Az NR tesztek reliabilitásmutatói a tesztek belső konzisztenciájából indultak ki, esetünkben a döntés konzisztenciája (decision-consistency) lehet a reliabilitás alapja. Mégpedig a leggyakrabban az, hogy mennyire konzisztensen osztja két csoportba a teszt azokat, akik a tananyagot már elsajátították és akik még nem sajátították el. A másik megközelítés az általánosíthatóság-elméletnek (generalizibility theory, G theory; a következőkben G elmélet) a CR tesztekre adaptálásából fejlődött ki. Mindkét megközelítést csak röviden, néhány példán keresztül mutatom be. A szemléletmódra helyezve a hangsúlyt, a matematikai háttérre csak az egyszerű számításokon alapuló, szemléletesen interpretálható reliabilitásmutatóknál térek ki. Természetesen mindegyik területen vannak kifinomult, csak számítógépen kivitelezhető elemzési eljárások is.

A döntési konzisztencián alapuló reliabilitásmutatók legegyszerűbb esetben annak megbízhatóságát jellemzik, ahogy a tanulókat az elsajátította — nem sajátította el (master — nonmaster) osztályokba soroljuk. A megismételt vagy ekvivalens tesztváltozatokkal végzett párhuzamos tesztelés adatait felhasználva a konzisztensen minősített (consistently classified) tanulók arányát tekintjük a reliabilitás mértékének. Jelöljük azoknak a tanulóknak az arányát, akiket az első tesztelés alapján az "elsajátította", a második (megismételt vagy paralel) tesztelés során a "nem sajátította el" csoportba soroltunk a $p(+, -)$ jellel, a mindkét teszt alapján az "elsajátította" csoportba soroltak arányát a $p(+ +)$ jellel stb. Ekkor a konzisztensen minősítettek aránya, amit $p(0)$-lal jelölünk, a következő lesz:

$$p(0) = p(+ +) + p(- -).$$

A számítást kiterjeszthetjük arra az esetre, amikor nemcsak két, hanem tetszőleges m számú teljesítményosztályunk van. Ekkor ugyancsak azok arányát

kell összeszámolnunk, akiket az első és második mérés azonos osztályba sorolt, tehát:

$$p(0) = \sum p(kk),$$

ahol p(kk) azok arányát jelöli, akiket mindkét tesztelés a k-adik kategóriába sorolt.

Könnyen belátható, hogy ez a megoldás még egy különösen jól működő tanítási módszer végén alkalmazott záróteszt reliabilitásának kiszámítására is alkalmas, ahol esetleg a tanulók 85—90%-a elérte az elsajátítás kritériumát. Itt a kis szórás miatt a korrelációs technikán alapuló számítás félrevezető eredményeket szolgáltatna. A p(0) emellett könnyen interpretálható, a 100p(0) azt mutatja meg, hogy a teszt használatával a tanulók hány százalékát osztályozzuk (minősítjük) megfelelően (Swaminathan et al., 1974).

A fenti formula gyengéje viszont az, hogy (a véletlen egybeesések miatt) p(0) akkor is 0-nál nagyobb értéknek adódik, ha a vizsgált tanulócsoport tagjait egymástól függetlenül, véletlenszerűen soroljuk osztályokba. A tesztadatok alapján adhatunk egy becslést a puszta véletlen alapján konzisztensen osztályba soroltak arányára, ez pedig általános esetben, m osztályra a következő:

$$p(c) = \sum p(k.)p(.k),$$

ahol p(c) a véletlenszerűen konzisztens osztályba sorolások aránya, p(k.) és p(.k) pedig azoknak az aránya, akiket az első, illetve a második teszt a k-adik osztályba sorolt. A véletlenszerűen konzisztens osztályba sorolás korrekciójára a szerzők Cohen (1960) által nominális skálák közötti megegyezés mértékének jellemzésére kidolgozott formulát (Cohen-féle kappa) javasolják:

$$\kappa = \frac{p(0) - p(c)}{1 - p(c)}.$$

A Cohen-féle kappa már rendelkezik azzal a tulajdonsággal, hogy értéke 0 és +1 között változik. Akkor 0, ha az osztályozás konzisztenciája a véletlenszerűen várható értékkel egyezik meg. Negatív csak abban az abszurd esetben lehet, ha az osztályba sorolás konzisztenciája rosszabb, mint amit a puszta véletlen alapján várhatunk. Akkor lesz +1, ha a két teszt alapján végzett osztályba sorolás pontosan megegyezik.

A p(0), illetve a kappa koefficiens kiszámítása és értelmezése rendkívül egyszerű, ezzel szemben eléggé körülményes a hozzá szükséges adatok megszer-

zése, vagyis a kétszeri tesztelés. Valószínűségelméleti tételek felhasználásával (és gyakorlatilag bonyolult számítások árán) azonban ezeket a mutatókat egyetlen tesztelés eredményei alapján is lehet becsülni. Mivel az adatok feldolgozása egyébként is számítógéppel történik, a bonyolult számítások nem jelentenek lényeges korlátot. Különböző megfontolások alapján különböző becslő formulákhoz juthatunk, bonyolultságuk miatt azonban ezekkel nem foglalkozunk. (Ismertetésüket ld.: Subkoviak, 1980).

A CR tesztek megbízhatóságának elemzésére egyre gyakrabban használják az általánosíthatóság-elmélet eszközeit. Magát az elméletet nem a CR tesztelés problémáinak megoldására dolgozták ki, hanem az NR tesztek elméleteit kívánták kiterjeszteni. A megoldás azonban olyan átfogónak bizonyult, hogy a CR tesztek speciális eseteire is alkalmazható. A kiindulást Cronbach és munkatársainak írása (1963) jelenti, melyben a klasszikus tesztelmélet kötöttségeinek feloldására, a reliabilitásmutatók kiszámíthatóságához megkövetelt feltételek liberalizálására javasoltak elemzési modelleket. A klasszikus elmélet alapvető problémája ugyanis az, hogy a legtöbb mutató kiszámításához paralel tesztelésre van szükség, ami a gyakorlatban precízen soha nem kivitelezhető. Az általánosíthatóság-elmélet alapgondolata ennek megfelelően az, hogy paralel mérések helyett tetszőleges (a modell keretein belül mozgó) mérés megfelelő a számítások elvégzéséhez, ha a két mérés feltételeinek a különbözőségét is matematikailag kezelhetővé tesszük. A klasszikus tesztelmélet feltételezi, hogy a mérés hibája egy meghatározott forrásból származik és meghatározott eloszlást mutat, a G elmélet alkalmas a különböző forrásokból származó hibák kezelésére is.

A G elmélet alapfeltevése szerint létezik mind a vizsgált személyeknek, mind a tesztitemeknek egy-egy univerzuma, melyből az aktuális méréshez mind a személyeket, mind pedig az itemeket véletlenszerűen választottuk ki. Ekkor egy p személynek az i item megoldásával elért X(p,i) pontszámának várható értékét a következő lineáris modellel adhatjuk meg:

$$X(p,i) = m + m(p) + m(i) + m(p,i) + e,$$

ahol m a populációba tartozó összes személynek az itemek univerzumára számított teljes átlaga, m(p) az egyén hatása a pontszámra, m(i) az item hatása a pontszámra, m(p,i) az egyén és az item interakciójának hatása a pontszámra, e pedig a mérési hiba.

Hasonlóképpen értelmezhetjük a mérési feltételek egy univerzumát, és a modellt kibővíthetjük az X(p,i,j) pontszámra ható feltételeket reprezentáló m(j) és a kölcsönhatásokat reprezentáló m(p,j), m(i,j) és m(i,p,j) tagokkal.

A feltételeken belül már sokféle speciális mérési helyzetet vehetünk figyelembe, például a két mérés között eltelt időt, a fejlesztés, speciális kezelés vagy oktatás hatását (van der Kamp, 1976). Az eljárás nevében szereplő "általánosíthatóság" arra utal, hogy olyan kérdésekre keresi a választ, mint például: az itemek univerzumából véletlenszerűen kiválasztott halmazzal mint teszttel elvégezve a mérést, ennek eredményeit mennyire általánosíthatjuk arra az esetre, ha egy másik mintával végezzük a mérést, ha a feltételek mások stb.

A modell kezelésére a varianciaanalízis megfelelő matematikai eszköznek bizonyul, melynek segítségével ki lehet számítani a varianciának a különböző forrásokból eredő komponenseit. A CR mérésnek azt az alapelvét, hogy a vizsgált egyéneket nem egymáshoz akarjuk viszonyítani, hanem egy, a csoport eredményeitől független kritériumhoz, az eljárás úgy veszi figyelembe, hogy a hibavariancia számításakor a (négyzetes) eltéréseket nem a csoportátlagtól, hanem egy rögzített referenciaponttól számítja. Például egy olyan tesztnél, amellyel az "elsajátította — nem sajátította el" minősítést kívánjuk elvégezni, a referenciapont az elsajátítás kritériumaként megadott ponthatár (cut-off score) lehet.

A varianciákat felhasználva a reliabilitásmutatókhoz hasonló indexeket lehet készíteni. Gyakoribb azonban a különböző függőségi (dependability) indexek meghatározása. Ezek segítségével megítélhetjük, hogy a teszteredmények mennyire függetlenek valójában azoktól a feltételektől, amelyektől feltevéseink szerint nem szabad függeniük. A G elmélet alkalmazását inkább jellemzi az eredmények többoldalú mérlegelése, mint az egyetlen idexre alapozott döntés (Brennan, 1980; Cardinet—Tourneur—Allal, 1976).

Mérleg: megoldott és megoldatlan problémák

Összegzésként tekintsük át, mivel járult hozzá világszerte a kritérium-orientált mérés elmélete és gyakorlata a pedagógiai értékelés eszköztárához, mivel járulhat hozzá a magyarországi fejlődéshez, és mi az, amivel egyelőre vagy véglegesen ez a megközelítés is adós marad.

Mérleget készítve, a nyereség oldalán számos technikai részletkérdés megoldását említhetnénk. E részmegoldások jelentőségét azonban nem szabad alábecsülnünk: többnyire éppen azokat az ellentmondásokat oldották fel, amelyek legjobban szorították a pszichológia klasszikus tesztelméletének keretei közé szorított pedagógiai mérési gyakorlatot. Az újszerű megoldások sajátos szemléletmóddá állnak össze, kialakulóban van egy többé-kevésbé egysé-

ges tesztkészítési metodika, és a tesztek sokféle funkciójához alkalmazkodó sokszínű adatelemző eszköztár. A fejlődés a mindennapos pedagógiai gyakorlatban használható értékelőeszközök terén a leglátványosabb. A gyorsan változó, változtatható-fejleszthető pszichikus tulajdonságok állnak a kritérium-orientált tesztelés fókuszában, míg a klasszikus tesztelmélet érdeklődése a személyiség legáltalánosabb, stabil vagy lassan változó tulajdonságainak (intelligencia, kreativitás, temperamentum, introverzió-extroverzió, maszkulin-feminin jelleg stb.) mérése körül koncentrálódott.

A hiány oldalra tekintve legszembetűnőbb, hogy a nagy szintézis (legalábbis egyelőre) még várat magára. Egyelőre nem látszanak egy egységes matematikai elmélet körvonalai, a klasszikus tesztelmélet néhány egyszerű alapfeltevése helyére különböző feltételrendszerek léptek. Nem hozott megoldást a CR mozgalom a pedagógiai-pszichológiai értékelés legizgalmasabb problémáira, nem enyhítette a személyiség legáltalánosabb tulajdonságainak mérése körüli elméleti és gyakorlati bizonytalanságokat. Nem tudunk kritérium-orientált teszteket készíteni az intelligencia mérésére, mint ahogy a kreativitástesztek a priori validálásának is megvannak a maga korlátai. Nem küszöböli ki a CR tesztelés a szakértői kompetencia szerepét, sőt éppen felértékeli azt. Egy intelligenciatesztet jórészt technikai úton ki lehet kísérletezni, statisztikai eszközökkel ki lehet szűrni a "g faktort" legjobban mérő itemeket, de nélkülözhetetlen a szakértői kompetencia például a gáztörvényekről tanultak vizsgálatára szolgáló teszt kifejlesztéséhez. Módszereket találunk viszont a CR értékelés eszköztárában a kompetencia kontrollálására, kompetenciák összemérésére.

A magyarországi helyzetet és perspektívákat illetően csak bízhatunk a paradoxonok mielőbbi feloldódásában. Kevésbé problematikusnak látszik a létező tesztkészítő technikákhoz a megfelelő adatfeldolgozó, -elemző eljárások adaptálása. Sokkal súlyosabbnak tűnik az az ellentmondás, mely szerint egyrészt a CR tesztek készítésének technikája, a világszerte meginduló mozgásokkal egyidőben, bizonyos mozzanatait tekintve időben és színvonalban is azt megelőzve megjelent, ugyanakkor a mindennapos gyakorlatot, a tesztek készítésének és használatának kultúráját aligha lehetne világszínvonalúnak nevezni.

IRODALOM

Báthory Zoltán (1985): Tanítás és tanulás. Tankönyvkiadó, Budapest.

Berk, R. A. (Ed. 1980): Criterion-referenced measurement: The state of the art. The Johns Hopkins Press Ltd., London.

Brennan, R. L. (1980): Applications of generalizibility theory. In: Berk, R. A.: i. m.

Cardinet, J.—Tourneur, Y.—Allal, L. (1976): The generalizibility of surveys of educational outcomes. In: van der Kamp, L. J. Th. ed.: Advances in psychological and educational measurement. Wiley, London.

Cohen, J. (1960): A coefficient of agreement with provision for nominal scales. = Educational and Psychological Measurement 20., 37—46. p.

Cronbach, L. J.—Rajaratman, N.—Gleser, G. C. (1963): Theory of generalizibility: A liberalization of reliability theory = British Journal of Statistical Psychology, 16., 137—63. p.

Csáki Imre—Nagy József (1976): Alsó tagozatos szöveges feladatbank = Acta Univ. Szeg. de A. J. nom. Sectio Paed. Ser. Spec., Szeged.

Fricke, R. (1974): Kriteriumsorientierte Leistungsmessung. Verlag W. Kohlhammer, Stuttgart.

Glaser, R. (1963): Instructional technology and the measurement of learning outcomes: Some questions = American Psychologist, 18., 519—21. p.

Hambleton, R. K. (1980): Test score validity and standard-setting methods. In: Berk, R. A.: i. m.

Hambleton, R. K.—Novick, M. R. (1973): Toward an integration of theory and method for criterion-referenced tests = Journal of Educational Measurement, 10., 159—70. p.

Hively, W. (ed. 1974): Domain-referenced testing. Educational Technology Publications, Englewood Cliffs, New Jersey.

Horváth György (1985): Tesztelmélet: problémák és perspektívák = Pszichológia 1. sz. 53—78. l.

Kamp, L. J. Th. van der (1976): Generalizibility and educational measurement. In: van der Kamp, L. J. Th. Ed.: Advances in psychological and educational measurement. Wiley, London.

Kiss Árpád (1960—61): Iskolai tanulóink tudásszintjének vizsgálata 1—3. = Pedagógiai Szemle, 1960. 3. sz. 194—206. l., 7/8. sz. 585—593. l., 9. sz. 775—784. l.; 4. 1961, 7/8. sz. 600—613. l.

Kiss Árpád (1961): Docimológia, osztályozás, mérés = Pszichológiai tanulmányok 3. Akadémiai Kiadó, Budapest, 253—266. l.

Lord, F. M.—Novick, M. R. (1968): Statistical theories of mental test scores. Reading, Mass., Addison—Wesley.

Nagy József (1972): A témazáró tudásszintmérés gyakorlati kérdései. Tankönyvkiadó, Budapest.

Nagy József (1975): A témazáró tesztek reliabilitása és validitása (STT 18. kötet) = Acta Univ. Szeg. de A. J. nom. Sectio Paed. Ser. Spec., Szeged.

Nagy József (1973): A standard osztályzat = Pedagógiai Szemle, 3. sz. 225—234. l.

Popham, W. (Ed. 1971a): Criterion-referenced measurement, An introduction, Englewood Cliffs, N. J.: Educational Technology Publications.

Popham, W. (1971b): Educational criterion measures. New York, American Book.

Popham, W. J. (1978): Criterion-referenced measurement. Englewood Cliffs, N. J. Prentice Hall.

Popham, W. J.—Husek, T. R. (1969): Implications of criterion-referenced measurement. = Journal of Educational Measurement, 6., 1—9. p.

Strittmatter, P. (ed. 1973): Lernzielorientierte Leistungsmessung. Weinheim, Beltz.

Subkoviak, M. J. (1980): Decision-consistency approach. In: Berk, R. A.: i. m.

Swaminathan, H.—Humbleton, R. K.—Algina, J. (1974): Reliability of criterion-referenced tests: A decision-theoretic formulation = Journal of Educational Measurement, 11. No. 4., 263—67. p.

Ward, J. (1970): On the concept of criterion-referenced measurement = British Journal of Educational Psychology, 40., 314—23. p.