

SZEMLE

HORVÁTH GYÖRGY

Műszaki Egyetem, Budapest

BESZÉLGETÉSEK A MODERN TESZTELMELETRŐL

BEVEZETÉS

A modern tesztelmélet legalább negyedszázados múltra tekint vissza. A modern jelzõt mégis indokolttá teszi több körülmény: a klasszikus tesztelmélettől való elkülönítés igénye; a modern számítógépi lehetőségekhez való szoros kapcsolódás; és az a tény, hogy a pszichometria egyik ma is бурjánzóan fejlődő ágazatáról van szó, amit az e témakörben megjelenő könyvek, tanulmányok évről-évre növekvő száma is jelez. Nehéz volna felbecsülni, hogy hány tételből is állna egy többé-kevésbé teljes bibliográfia, de biztos, hogy ma már tekintélyes kötetet tenne ki csak a modern tesztelmélettel foglalkozó munkák címjegyzéke is. A tesztelmélet legtekintélyesebb folyóirat-műhelyének, a *Psychometrikának* huszonöt évfolyamát áttekintve Charles Lewis a folyóiratban megjelenő közlemények százalékarányainak alakulásával illusztrálta azt a fellendülést, amit a tesztelméleti kutatásokban az új, modernebb szemléletmód hozott (1. ábra).

Charles Lewis ábrája azt is érzékelteti, hogy a tesztelmélet klasszikus korszakának virágkora a harmincas-negyvenes évekre esett. A két legjellemzőbb problémakör akkoriban a faktoranalízis módszertana és értelmezése, valamint a klasszikus tesztelmélet alapfogalmainak (reliabilitás, validitás) a tisztázása volt. A hatvanas évekre az ez irányú kutatások ellanyhultak, kifulladásra; úgy tűnhetett, hogy a tesztelmélet fejlődése lezárult. Az összefoglaló munkák korszaka következett. Német nyelvterületen LIENERT (1961) kézikönyve reprezentálja leginkább a fejlődésnek ezt a szakaszát, hazánkban is ez a legismertebb tesztelméleti kézikönyv. A nagy számú angol nyelvű összefoglaló mű közül kiemelkedő szerepű, afféle

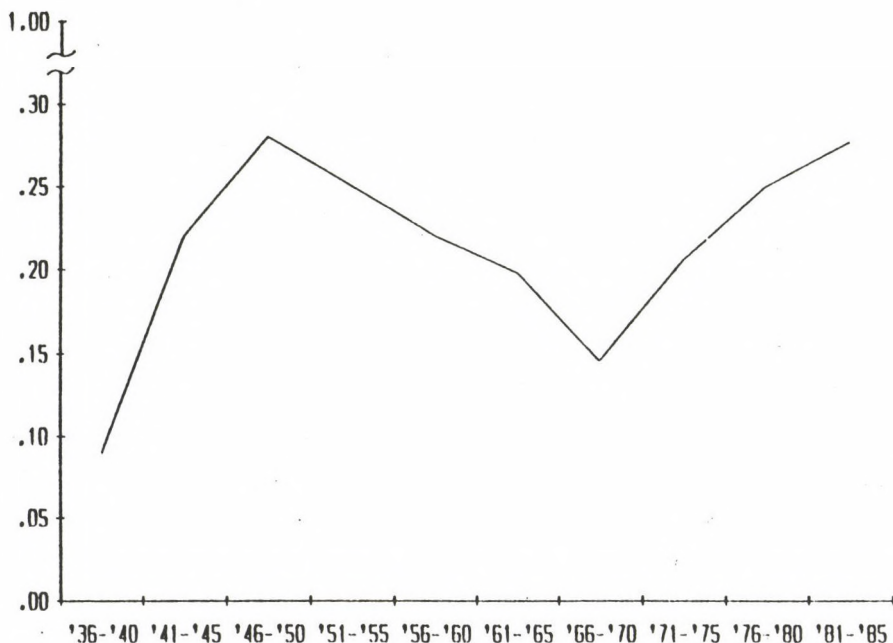
„mérőföldkőnek” számít H. GULLIKSEN (1960) munkája, mert a klasszikus tesztelmélet matematikai szempontból is legszabatosabb, legkiérleltőbb, axiomatikus felépítését mutatja be.

Egy másik jelentős mű néhány évvel később született meg ugyanabban a princetoni műhelyben, amelyben Gulliksen is dolgozott. A F. Lord és M. Novick nevéhez fűződő vaskos kötet már átmenetet képez a klasszikus tesztelmélet és a modernebb irányzatok közt. Felépítése olyan, hogy a „gyenge előfeltevésekre” épülő tizenkilenc klasszikus tesztelméleti fejezethez szervesen illeszthető volt a „szigorú előfeltevésekre” alapozott

1. ábra

Ch. Lewis (1983) grafikonja a Psychometrika c. folyóiratban megjelenő tesztelméleti publikációk számarányának alakulásáról.

(A vízszintes tengelyről a publikáció évét olvashatjuk le, a függőleges tengely a tesztelméleti tárgyú cikkek arányát mutatja.)



modelleket bemutató öt fejezet, amely a haditengerészet számára dolgozó Allan Birnbaumnak az elméleti eredményeit publikálta.

Közben, ahogyan az lenni szokott, másutt is megjelentek az újnak a csírái. Koppenhágában egy magányosan dolgozó dán statisztikus, G. RASCH egészen eredeti kérdésfeltevésből kiindulva jutott el az intelligenciamérés egy új matematikai modelljéhez, amelynek három változatát bemutató könyve ugyanabban az évben jelent meg (1960), mint Gulliksen klasszikus monográfiája.

A klasszikus tesztelmélet és modern tesztelmélet közti korszakhatár tehát 1960 körülre tehető. Igaz, a legtekintélyesebb mai angol pszichometrikus, H. Goldstein (GOLDSTEIN, WOOD, 1989) összefoglaló tanulmányában az új modellek történetének „ötven esztendejét” emlegeti. Egyrészt, hogy hangsúlyozza a klasszikus tesztelmélettel való folytonosságot, másrészt, hogy kiemelve az angol prioritást, az edinburgh-i tesztpszichológusok (G. Ferguson, D. W. Lawley, D. J. Finney) úttörő szerepét.

A modern tesztelmélet valóban nem veti el mindenestől a klasszikus eredményeket, bár más, szilárdabb elméleti alapokra szeretné helyezni a mentális mérések módszertanát. A modern tesztelmélet és a klasszikus tesztelmélet viszonyát, és egyúttal a klasszikus tesztelmélet gyenge pontjait is a legmélyebben és legegységesebben talán Gerhard H. FISCHER 1974-ben megjelent monográfiája világította meg, mely a német nyelvterületen igen nagy szerepet játszott az új elmélet és gyakorlat meggyökeresedésében. A klasszikus tesztelmélet kritikájának leglényegesebb pontja a legfontosabb tesztmutatók „populációfüggőségére” (mintavétel-függőségére) irányul. A tesztmutatók értéke függ a kiszámításuk alapjául szolgáló mintapopulációtól, ezért eltérő populációkból származó egyének teszt-eredményei általában nem objektíven összehasonlíthatók, aminek számos gyakorlati következménye közül az egyik legsúlyosabb a változásmérések problematikussága.

Rasch kérdésfeltevése éppen az objektivitás feltételeire vonatkozott. A tudományos mérés általános követelményeit a tesztmérésekre alkalmazva fogalmazta meg az összehasonlítások „speciális objektivitásának” követelményét, amely Rasch szerint azt jelenti, hogy két „objektum” (például személyek) összehasonlításakor az eredmény nem függhet attól, hogy az összehasonlításhoz felhasználható „ágensek” (például itemek) közül éppen melyiket alkalmaztuk. Ez a követelmény a klasszikus tesztelméletet jellemző „populációfüggőség” meghaladását igényli: a képességek mennyiségi összehasonlításának eredménye nem függhet attól, hogy mely ite-

mekkel végzik, és az itemek egymáshoz viszonyított nehézsége sem függhet attól, hogy mely személyek eredményeit használták fel becslésükre; az itemek kalibrálása tehát „populációfüggetlen”, „mintavételtől független” (sample-free) legyen.

Benjamin Wright, a Rasch-modell legtekintélyesebb amerikai szakembere számos tanulmányban foglalkozott a méréselméleti alapokkal is, és arra mutatott rá, hogy a Rasch-modell specifikus objektivitása révén olyan követelmények valósulnak meg — a mérési eljárás objektivitása és az eredmények additivitása —, amelyeket már L. L. Thurstone, majd Luce és Tukey is megfogalmazott, de amelyeket a tesztmérésekben csak a Rasch-modell tesz lehetővé. Ebben az értelemben a Rasch-modell a klasszikus tesztelméletben is felmerült igényeket elégíti ki.

Kétségtelen az is, hogy az újabb modellek egyes fogalmi elemei a klasszikus vizsgálatok keretében alakultak ki. A modern tesztelmélet nem az összesített tesztmutatókra, hanem az egyes itemek tulajdonságaira helyezi a hangsúlyt. De ez a szemlélet jelentkezett már W. RICHARDSON-nak (1936) abban a törekvésében is, hogy a leglényegesebb tesztmutatókat (a reliabilitást és a validitást) itemmutatókra vezesse vissza. A Goldstein (1989) által úttörőként méltatott edinburgh-i pszichológusok ugyancsak az itemtulajdonságok matematikai jellemzésére törekedtek. D. J. Finney a gyógyszerhatás-vizsgálatokra kidolgozott probitanálízis fogalmait alkalmazta a tesztstúdióra: a képességet dózisként, az itemet organizmusként kezelte. Lawley Richardsonhoz hasonlóan megpróbálta a klasszikus tesztmutatókat itemparaméterekkel kifejezni. Az ugyancsak itemtulajdonságokat tanulmányozó G. Ferguson volt az, aki már 1942-ben azt fontolgatta, hogy egy itemre valamely személy által adott helyes válasz valószínűsége a normális eloszlásfüggvény görbéjével jellemezhető. Hasonló megállapításra jutott L. R. TUCKER (1946), aki a helyes válasz valószínűségét jellemző görbének a ma is leghasználatosabb nevét adta: item-jelleggörbe (item characteristic curve, ICC). De csak később, P. Lazarsfeld látensstruktúra-analízis modellje révén (LAZARSFELD, HENRY, 1968) vált tudatossá, hogy a jelleggörbe (amit Lazarsfeld nyomvonalnak nevezett), illetve a megoldás valószínűségét meghatározó függvény egyúttal egy-egy modellt is meghatároz. Lazarsfeld modellje azért is előfutára az új tesztelméleti modelleknek, mert valószínűségi jellegű: a látens struktúra analízise nem besorolta a személyeket, hanem csak azokat a besorolási valószínűségeket határozta meg, amelyekkel egy-egy válaszmintázat valamely látens osztályhoz kapcsolhatja a válaszolót. Bár az attitűdvizsgálatokból jól ismert Guttman-skála nem valószínűségi, hanem determinisztikus jellegű,

néhány sajátossága miatt GUTTMAN (1945) „perfect” skáláját is az újabb tesztmodellek előkészítői közé szokták sorolni. A Guttman-féle modellben két változó, az itemet jellemző „nehézség” és a személyt jellemző „beállítódás” egymást határozzák meg. Item és személy ugyanazon a skálán helyezkedik el, szerepük szimmetrikus, egymásra vonatkozásukban létesítik a skálát.

Az újabb tesztmodellekre általában igaz, hogy elsősorban az egyes itemek tulajdonságait, az itemekre adott helyes válaszok meghatározó tényezőit kívánják jellemezni. Ez fejeződik ki az USA-ban meghonosodott terminológiában is, amely a modern tesztelméletet Item Response Theory néven foglalja össze. (Az elnevezés F. Lordtól ered, Európában inkább csak az elmélet amerikai megjelenési formáira alkalmazzák. Európában viszont Rasch nyomán eléggé elterjedt a „probabilisztikus modellek” megjelölés is.) Minden modern tesztelmélet lényege az a valószínűségfüggvény, amely megmutatja, hogy miként változik egy-egy item megoldásának valószínűsége a személyeket és az itemeket közös skálán jellemző paraméterértékek változásakor. Az egyes modellek abban különböznek egymástól, hogy a helyes megoldás valószínűsége és az item-, illetve személyparaméterek közt milyen függvénykapcsolatot (például normális eloszlásfüggvényt, logisztikus függvényt vagy más monoton függvényt) határoznak meg.

Rasch egy matematikailag könnyen kezelhető függvényre építette modelljét, Lord modellfüggvénye a normális eloszlásfüggvény. (A két függvény grafikonja egyébként majdnem teljesen egybevágh, mindkettő elnyújtott S alakú, de matematikai leírásuk igen eltérő.) Rasch modelljében két paraméter (képességparaméter és nehézségparaméter) található, Birnbaum logisztikus modelljében egy képességparaméter és három itemparaméter (nehézségmutató, elkülönítésmutató, kitalálhatóságmutató) szerepel.

Az első probabilisztikus tesztmodellek (Rasch, Lord, Birnbaum) óta folyvást újabbak és újabbak születnek. Némelyek megőrzik a logisztikus (vagy normális eloszlás-) függvényt, de kissé módosítják a paramétereket. Rasch logisztikus modellje például első változatában dichotom volt (csak helyes/nem helyes válaszlehetőséget ismert), a modell általánosabb formája polytom (több válaszfokozatot is elfogad). A paraméterek számának növelése újabb és újabb elméleti modelleket terem; egy másik, az alkalmazhatóság szempontjából termékenyebb fejlesztési lehetőség pedig az egyes paraméterek (például a nehézségparaméter) összetevőkre való felbontása. Az újabb modellek keletkezésének tág teret ad az egy-egy konk-

rét kérdésfeltevéshez vagy adatstruktúrához jobban illeszkedő modellfüggvények keresése is.

Charles LEWIS szerint (1986) a különböző tesztmodellek száma 1986-ban ötven körül lehetett. KUBINGER egy rendszerező áttekintésben (1988) a legismertebb újabb modellek közül a következőket sorolja fel: Fischer lineáris logisztikus modellje (LLTM), Bradley, Terry és Luce (BTL-) modellje, Sixtl „felfejtő” modellje, Masters „Partial Credit” modellje, Spada logisztikus automata modellje, Whitely sokdimenziós modellje, Rasch Poisson-modellje, Scheiblechner latenciamodellje, Kempf dinamikus tesztmodellje, White három személyparaméterű modellje. És Kubinger áttekintésében még nem is szerepelnek olyan jól ismert és fontos modellek, mint a Mokken-skála vagy az ún. „rating scale” modellek gyorsan bővülő családja (WRIGHT, MASTERS, 1982; FISCHER, PARZER, 1991) vagy a látensstruktúra-analízissel egybeötvöződő ún. hibrid modellek (FORMANN, 1984; YAMAMOTO, 1989).

A különböző tesztmodellek nehezen áttekinthető sokaságában nemcsak történelmi okokból van kitüntetett szerepe a Rasch-modellnek. Nem is csupán azért, mert az említett modellek jelentős része is a logisztikus Rasch-modell általánosítása vagy speciális esete, hanem főként azért, mert a Rasch-modell különlegesen kedvező matematikai tulajdonságokkal rendelkezik. Ennek lényege a modellekben előforduló képességparamétereknek és az itemparamétereknek egymástól való szeparálhatósága a becslési folyamat során: az ún. feltételes likelihood eljárás alkalmazásával kiszámíthatók a képességparaméterek anélkül, hogy az itemparaméterek ismeretére szükség volna; a képességparaméterek kiszámításához tehát elégséges az itemmegoldások számának (a megoldásmátrix sorösszegeinek) ismerete. Ugyanígy az itemparaméterek meghatározásához sem szükséges a képességparaméterek ismerete, itemenként a személyek helyes megoldásainak száma (a megoldásmátrix sorösszegeinek ismerete), vagyis a nyerspontérték elégséges a képességparaméterek becsléséhez. A peremösszegek tehát tartalmazzák a teljes megoldásmátrixból mindazt az információt, ami a paraméterek (szeparált) kiszámításához elégséges, a peremösszegek „elégséges statisztikát” alkotnak. Az elégséges statisztika fogalmát még a húszas években R. A. FISHER (1959), a kitűnő angos statisztikus vezette be, és már Rasch is hangsúlyozta, hogy a paraméterek szeparálhatósága révén „a specifikus objektivitás fogalma ténylegesen szorosan kapcsolódik R. A. Fisher elégségesség-fogalmához” (RASCH, 1977, 66. o.).

Ez a figyelemre méltó tulajdonság, vagyis a peremösszegek elégséges statisztika mivolta, az IRT-modellek közül csakis a Rasch-modellre (és

közvetlen „rokonaira”) igaz. FISCHER (1974) ugyanis bebizonyította, hogy ha a személyek nyerspontértéke elégséges statisztika és fennállnak az IRT-modellekben általában megkövetelt feltételek, akkor e feltevésekből szükségképpen következik, hogy az így meghatározott modell Rasch-modell lesz.

A Rasch-modell specifikus objektivitása, a peremösszegek elégséges statisztika mivolta és a modell kétféle paraméterének (képesség- és item-paramétereknek) a szeparálhatósága szorosan összekapcsolódó fogalmak. A paraméterek szeparálhatósága teszi lehetővé a paraméterértékek becslését ún. feltételes maximum likelihood eljárással. Ezt E. ANDERSEN (1970) mutatta ki, de sok munkába és próbálkozásba került, amíg Andersen algoritmusát alapján (FISCHER, 1974), majd más algoritmusok alapján a gyakorlatban is használható (WRIGHT, PANCHAPAKESAN, 1969; WINGERSKY, 1983) számítógépi programokat sikerült szerkeszteni. Ma is intenzív munkálatok folynak egyre szélesebb körben alkalmazható programok kidolgozására (FISCHER, PARZER, 1991; LUNZ, WRIGHT, LINACRE, 1990; MISLEVY, STOCKING, 1989).

A modellek alkalmazásának egyik gyakorlati problémája, hogy ráillik-e egy-egy modell a feldolgozandó adatokra. A modern tesztelmélet hívei e modellek egyik legfontosabb előnyének éppen azt tartják, hogy érvényességük (például a „populációfüggetlenség”) egyértelműen ellenőrizhető. Már RASCH (1960) is bemutatott erre egyszerű módszereket, E. ANDERSEN (1973) pedig megmutatta, hogy a modell illeszkedése a likelihood-arány próba egy változatával is ellenőrizhető. Több más eljárás is született más IRT-modellek ellenőrzésére is (HAMBLETON, SWAMI-NATHAN, 1985), ez is mozgásban lévő kutatási terület. A modellilleszkedés sérelmeinek kimutatása sem okvetlenül negatív eredmény a modern tesztelméletben, amely éppen ilyen módon teszi lehetővé az itemek kultúrafüggőségének vagy más szisztematikus torzításoknak (bias), az itemek eltérő működésének objektív tanulmányozását. Az ilyen célú kutatások a modern tesztelmélet alkalmazásának ugyancsak kedvelt terepét alkotják (HULIN, DRASGOW, 1983; HAMBLETON, ROGERS, 1986; TATSUKO, 1990). Az alkalmazások egy része a klasszikus tesztelméletben is felmerült feladatok eredményesebb megoldását célozza. A legfontosabb ilyen feladatkör az itemek nehézségi szintjének összemérése (COOK, EIGNOR, 1989), a tesztek szintegyeztetése (equating), amely más oldalról a változásmérések objektív megoldásának is feltétele (FISCHER, PARZER, 1991). Csakis az IRT-modellek alapján lehetett azonban reális célkitűzéssé az olyan adaptív tesztelés, ahol egy tesztmérés folyamatában

az itemek nehézségi szint szerinti adagolását az egyes egyének képesség-szintjéhez igazítják (LORD, 1970; FISCHER, PENDL, 1980).

Ha befejezésül azt próbáljuk meg címszavakban felvázolni, hogy melyek a modern tesztelmélethez kapcsolódó szerteágazó kutatások legújabb tendenciái, három kérdéskört említhetünk meg: a többkategóriás többdimenziós modellek új alkalmazási formái; a számítógépek közvetlen felhasználása a tesztelés folyamatában; és a közeledés a kognitív pszichológiához. Ez utóbbi törekvés különösen nagy jelentőségűnek tűnik, s nem csupán a kognitív pszichológia számára, hanem a tesztelmélet fejlődése szempontjából is. Hiszen nem nagyon túloz R. Mislevy aforisztikus megállapítása, amellyel az újabb útkeresést indokolja: „A mai tesztelmélet a huszadik századi statisztika alkalmazása a tizenkilencedik századi pszichológiára” (MISLEVY, 1989, 1.).

Beszélgetőpartnereink úttörői voltak és ma is főszereplői annak a pezsgő fejlődésnek, amelyet a fentiekben dióhéjban jellemeztünk. Pályafutásuk sok párhuzamosságot mutat, de néhány jellegzetes eltérés is hangsúlyt kap. Beszámolójuk, véleményük segít tájékozódni annak a folyamatnak a történetében, jelenében és távlataiban, amellyel a hazai pszichológia fejlődése eddig, érthető okokból, nem nagyon tudott lépést tartani. A felzárkózáshoz az út a tájékozódáson át vezet.

Ben Wright az egyik legtekintélyesebb amerikai egyetem, a főként postgraduate képzéssel foglalkozó Chicagói Egyetem professzora. Ez az intézmény, amely hatvanegy (!) Nobel-díjassal büszkélkedik, mindig a pszichometria egyik fellegránának számított, itt működött Thurstone, Richardson és Guilford is. 1961-ben itt volt vendégtanár G. Rasch, és ettől kezdve Ben Wright az új módszerek nagyhatású amerikai propagátorává és továbbfejlesztőjévé vált.

Charles Lewis ma azt a kutatási osztályt vezeti az Educational Testing Service princetoni központjában, amelynek élén nyugdíjba vonulásáig Fred Lord állt. Az ETS több mint kétezer dolgozót foglalkoztató intézmény, amely évente több százezer tesztvizsgálatot bonyolít le. Ez a „non-profit” szervezet 300 millió dolláros forgalmat bonyolít le évente, és megengedheti magának, hogy a szakma legjobbjait foglalkoztassa. Nincs a világon még egy hely, amely olyan tapasztalattömeg és olyan anyagi eszközök birtokában foglalkozna a tesztfejlesztés problémáival, mint az ETS.

Gerhard H. Fischer, a bécsi egyetem pszichológiai intézetének professzora több évtizede Európa legjelentősebb tesztelméleti szakértőjének

számít. Igényessége és sokoldalúsága iskolát teremtett, tanítványai sok ország tanszékein vannak jelen. Néhány éve Budapesten is járt, és több előadást tartott.

Az interjúk 1991. januárban, illetve 1992. januárban készültek.

Ben WRIGHT: MÓDSZERTANI FORDULAT KÖZELEDIK

— A Rasch-modell ma már a pszichológiában és azon túl is jól ismert fogalom, Rasch nevét jól ismeri a tudományos közvélemény. Rasch-ról, az emberről, személyiségéről azonban úgyszólván semmit sem tudunk. Önök barátok voltak, milyennek ismerte meg őt?

— Georg Rasch (aki 1901-ben született) szigorú, sőt talán rideg légkörű családban nevelkedett. Édesapja fiatal korában tengerészként, hajó-tisztként dolgozott, később matematikatanár lett és olyan önkéntes misszionárius, aki a matrózokat istenfélővé akarta nevelni. E puritán „keményfejű evangelista” harmadik, legkisebb fia, Georg Rasch azonban életvidám, tréfálkozásra hajló, derűs alkatú férfiú volt, aki nem vetette meg a földi jót: szeretett enni-enni, kedvelte a társaságot, a nagy ebédeket, a jó bort, a konyakot, a koktélokot... A testes, pirospozsgás ember szép kort, nyolcvan évet ért meg, és teljes életet élt. Tudósként az aprólékos-ságig gondos, a hibátlan eredményre törekvő szakember volt, kiváló matematikus, aki szerette, amit csinált, élvezte a munkáját. A diplomáját a koppenhágai egyetemen szerezte, de a fiatal matematikus pályafutásának jelentős állomása, amikor 1934-ben Rockefeller-ösztöndíjjal egy esztendő-t tölt a londoni Galton Laboratóriumban R. A. Fisher mellett. 1938-ban E. Rubin professzor kérésére tartott először statisztika előadásokat a koppenhágai egyetem pszichológiai tanszékén. Ez a munkakapcsolat azután több, mint harminc évig tartott.

— Miért olyan nagy jelentőségű Rasch munkássága a pszichometriai tesztelmélet újabb fejlődése számára?

— Miként azt bizonyára tudja, Georg Rasch munkássága az ötvenes években jutott döntő szakaszához, ekkor jutott el a fundamentális jelentőségű felismeréshez, hogy „probabilisztikusnak” nevezett tesztmodelljei alapján az itemek és a személyek jellemzői szeparálhatók. Vagyis ekkoriban fedezte fel, hogy megadható a személyek képességének olyan objektív becslése, amely független attól, hogy mely itemeket használtuk fel hozzá — tehát „tesztfüggetlen” —, és egyidejűleg elvégezhető a tesztek olyan

kalibrálása is, amely független attól, hogy mely személyek válaszoltak az itemekre, tehát az itemkalibrálás is független a mintapopulációtól.

Olyan kritériumok ezek, amelyeket már Edward Thorndike is megfogalmazott a húszas években, és amelyeket Louis Thurstone is meglehetősen részletesen fejtett ki 1925 és 1930 közt több tanulmányában, majd a harmincas években Guilford is megismételt, és közvetett formában Guttman is fenntartott a negyvenes években. De csak Thurstone-nak sikerült valamennyire is megközelítenie ezt a célkitűzést. Az általa alkalmazott módszerek azonban számos szempontból nem kielégítőek.

Georg Rasch teljesen önállóan és új utakon járva bukkant rá a megoldásra, amikor a személyek egy csoportjának különböző tesztekkel (és eltérő időpontokban) regisztrált teljesítményeit kellett összehasonlítani. Maga is csak fokozatosan ébredt rá — ez már az ötvenes évek vége felé volt —, hogy felfedezése milyen fundamentális jelentőségű.

— Ön az elsők közt ismerte fel Rasch művének újszerűségét és jelentőségét. Hogyan találkozott Ön Rasch elgondolásaival?

— Valóban, Dánia határain kívül én voltam az első kutató, aki egyáltalán valamilyen érdeklődést tanúsított Georg Rasch munkássága iránt. — A másik Gerhardt Fischer volt, a szomszédságukból, a bécsi egyetemről, aki 1966-ban ismerkedett meg Raschsal.

Amikor 1960 tavaszán Rasch a Chicagói Egyetemre jött, hogy új felfedezéseiről előadjon, már tisztában volt felfedezésének jelentőségével. Első előadásait nagy érdeklődés kísérte, sokan jöttek el meghallgatni. Előadásmódja azonban a társadalomtudósok számára túl komplikáltnak, túlságosan matematikainak bizonyult, a statisztikus szakembereket pedig sértette az a nem túlzottan diplomatikus bíráló, amit a szokványos statisztikai módszerekről elmondott. Így azután igen hamar megfogyatkozott a hallgatóság, és a heti két előadásának egyetlen egy hallgatója maradt: én. Rasch azonban ugyanúgy folytatta előadássorozatát, ugyanúgy tartotta meg óráit, mintha zsúfolt terem számára beszélne. Az óra végeztével azután együtt ebédeltünk és tovább beszélgettünk. Az a két dolog, ami Georgban engem ekkoriban megfogott és előadásaira vonzott, az egyrészt gondolkodásának tisztasága és következtetessége volt, másrészt az a kurázsi, az a szokatlan eltökéltség és bátorság, amellyel meghirdetett előadássorozatát akkor is folytatta, amikor egyetlen ifjú tanársegéd kivételével mindenki faképnél hagyta őt.

Engem már legelső előadásai is meggyőztek az item- és személyparaméterek szeparálhatóságának jelentőségéről. Nem kételkedtem benne, hogy a társadalomtudományok további fejlődéséhez nélkülözhetetlen lesz az új eljárás mód, amit akkor már ki is próbáltam. Az általam felvett adatokat Georggal közösen elemeztük. Fáradtságos papír-ceruza számításokat végeztünk, hiszen számítógép ehhez a munkához még nem állt rendelkezésünkre. Nekem ugyan volt már némi számítógépes gyakorlatom: egyetemünk első, UNIV-1 komputeréhez írtam néhány faktoranalízis-programot, és sok faktoranalízist végeztem. Georg elgondolásai azonban termékenyebbek voltak, mint a faktoranalízis, és én ezt már akkoriban kezdem felismerni, abban a három hónapban, 1960 tavaszán. Egy ideig azonban még a régi vágányon haladt tovább a tevékenységem: Georg visszament Dániába, engem más feladatok foglalkoztattak, így aztán nem tértem vissza az általa tanítottakhoz egészen 1963-ig vagy 1964-ig.

— Mi hozta a tényleges fordulatot az Ön munkásságában? Hogyan lett Ön végül is a Rasch-modell egyik legismertebb alkalmazójává és szószólójává?

— Raschsal való újabb találkozásaim hatására. 1964 tavaszán sikerült néhány hónapra családotól Dániába látogatnom. Minden héten négy vagy öt délelőttöt Georg házában töltöttem, ahol dolgozószozájában egy tábla mellett vitattuk meg a mérésprobléma matematikai vonatkozásait. Nagyon élveztem ezeket az egyéni konzultációkat, nem utolsósorban azért, mert Georg rendkívül világosan fejtette ki gondolatait, és hát — mint már mondtam — lebilincselő, jó humorú társalgó, rendkívül megnyerő egyéniség volt. Ugyanakkor olyan embernek ismertem meg, aki a maga útját járja a tudományban is. Úgy láttam, hogy az ő felfogása a statisztikai problémákról sokkal aktuálisabb és realistább, mint legtöbb nagynevű kollégámké, és úgy éreztem, hogy az a szemléletmód, ahogyan a mérésproblémákat megközelítette, gyökeresen más, mint amit ebben az országban és másutt is tipikusan műveltek.

Amikor tehát megtehettem, 1965 nyarán újból visszatértem Dániába, és ismét néhány hónapot töltöttünk közös munkával. Ez alkalommal egy tanítványomat is magammal hoztam, Bruce Chappmant, aki angol volt, és később Angliában lett az új módszerek egyik úttörője. Bruce és én naponta jártunk Georghoz tanulni, eszmét cserélni matematikai és méréselméleti problémákról.

Harmadszorra 1967 nyarán jártam Georgnál, amikor egy tengeri szigeten lévő nyaralójában töltöttünk egy hónapot. Ekkor is, szokás szerint,

nap mint nap részletekbe menő eszmecseréket folytattunk a bennünket foglalkoztató elméleti problémákról. Azt hiszem, nyugodtam állíthatom, hogy minden olyan jelentős újítás vagy felfedezés, amely a Rasch-moddal kapcsolatosan az elmúlt 23 évben megjelent, felmerült már ezekben a beszélgetéseinkben 1967 nyarán vagy még korábban. Persze ekkor mi ugyan-csak elméleti síkon mozogtunk, és így messzire előre szaladhattunk, anélkül, hogy a gyakorlati nehézségekkel is szembe kellett volna néznünk. A gyakorlati problémák megoldása, a technikai nehézségek leküzdése azonban még sok időt és nem kevés munkát igényelt.

— Hogyan jutott el a Rasch-modell az Egyesült Államokban a szélesebb körű gyakorlati alkalmazásig? Mik voltak a leküzdendő technikai és gyakorlati nehézségek?

— 1965 nyarán Chicagóban a Midwest Psychological Association szervezésében Bruce Chappmannal együtt szimpoziumot tartottunk a Rasch-módszerről. Ekkor mutattuk be az IBM-komputerhez írt Fortran-programunkat, amely a számítógép lehetőségeit igyekezett kihasználni Georg elgondolásainak gyakorlati alkalmazásához. Mindaddig egész Amerikában Bruce-on és rajtam kívül csak egy fizikusnő, N. Panchapakesan és két szakdolgozatot író egyetemi hallgató foglalkozott a Rasch-moddal, annak ellenére, hogy Rasch 1960-as könyvéről igen elismerő recenziót közölt egy tekintélyes pszichológus, Jane Loevinger. Rasch könyvére néhány helyen felfigyeltek, de a modell alkalmazásának semmi látható nyoma nem volt.

A számítógépi program azonban kimozdította az ügyet a holtpontról. Olyan algoritmust alkalmaztunk, amely némiképpen eltért a Rasch által használt ideális módszertől, de gyorsabban vezetett eredményre és majdnem olyan pontos volt. (Rasch eredetileg a feltételes valószínűségeken alapuló eljárást használta, mi viszont ún. nem feltételes algoritmust dolgoztunk ki.) 1966-ban publikáltuk is Fortran-programunkat, és még ma is csaknem mindenki ehhez a „nem feltételes” eljárásához folyamodik, ha Georg elgondolásait ténylegesen adatelemzésre alkalmazza.

A fejlődés ezután felgyorsult. 1967-ben az Educational Testing Service tudományos konferenciáján számoltam be az eredményekről, és ennek az előadásnak szélesebb körű visszhangja lett. Az American Educational Research Association szervezésében a módszer gyakorlati elterjesztésére workshopok sorozatát tartottuk éveken át, némelyiket Georg részvételével. A hétköznapi gyakorlatba a Rasch-modell akkor vonult be, amikor Portland állami iskoláinak a tesztvizsgáztatási rendszerét a Rasch-mo-

dellre építették, a mai napig tartó sikerrel. A másik terület, ahol a gyakorlati áttörés megindult, az egészségügyi-orvosi képesítés rendszere. A National Board of Medical Examiners 1977 óta a Rasch-modellre alapozza a tesztrendszerét. Ma már az alkalmazások rendkívül kiterjedtek, és néha egészen meglepő helyeken találkozunk velük.

Jómagam az elmúlt huszonöt évben kb. 45 Ph. D.-disszertáció elkészítését irányítottam, ezek több mint fele valamiképpen a Rasch-modellhez is kapcsolódott, és közülük mintegy tíz olyan volt, amely elméleti vagy módszertani vonatkozásban továbbfejlesztette a Rasch-méréseket. A legjelentősebb munkák szerzői alighanem David Andrich, Jeff Masters, Ronald Mead és Micheal Linackre voltak. Mindegyikük komolyan hozzájárult a mérési modellek fejlődéséhez, többen maguk is professzorok lettek, és már maguk is tanítványaik révén terjesztik a modern szemléletet, ami igen jó dolog.

— Lezárult-e a Rasch-modell elméletének és módszertanának a fejlődése vagy még számíthatunk új fejleményekre?

— Távol állunk még attól, hogy kitapasztaltuk volna a Rasch-modellben rejlő összes lehetőséget. Jelenleg Michael Linackre például igen termékeny irányban terjeszti ki a Rasch-modellt, amikor egy meglehetősen bonyolult szituáció értékelésére használja fel a modell egy újabb változatát.

Röviden arról van szó, hogy objektívebbé tudjuk tenni az olyan szituációkat, amikor például egy zsűri minden tagja egymástól függetlenül, megadott szempontok szerint pontozza egyes személyek különböző feladatokban nyújtott teljesítményét. Például, ha a vizsgabizottság szövettani metszetek alapján osztályozza a vizsgázókat, akkor kalibrálni tudjuk az egyes metszeteket, a szempontokat, amelyek szerint a patológusok osztályozzák a teljesítményt, és a zsűrit alkotó patológusokat is. Mindebből olyan mérést tudunk produkálni, amely a vizsgázó eredményét függetleníti attól, hogy melyik pontozóhoz kerültek a metszetei, vagy hogy éppen melyik metszetek alapján osztályozták, és így tovább. Ily módon a zsűri által végzett értékelés folyamatát nagyon kiegyensúlyozottá és korrektté tudjuk tenni. A zsűritagok statisztikai mutatói pedig lehetővé teszi a vizsgákat felügyelő bizottság számára, hogy ellenőrizze a pontozói teljesítményeket is, felfigyeljen a következetlenségekre és szükség esetén akár ki is zárja a további munkából a nem megfelelően pontozó bírakat, illetve vizsgázatókat.

Az az eljárás, amit erre a meglehetősen komplikált szituációra kidolgoztunk, szintén a Rasch-modellen alapul, jól bevált és sok területen alkalmazható. A University of Illinois például a fogyatékoság mértékét megállapító bizottságok munkájának összehangolására és objektívebbé tételére használja az általunk kidolgozott módszert. Linackre szeretne volna egy sport-világbajnokság pontozóbíráinak teljesítményét is értékelni ily módon, de az illetékes sportszövetség elzárkózott ez elől. Ők tudják, hogy miért.

— Akinek volt alkalma az Ön érdekes és lelkesült előadásait hallgatni, érezhette, hogy a „száraz” matematika miként kapcsolódik nagyon is emberi problémákhoz. Mégis meglepődve olvastam legújabb könyvét, amelynek már a címe is szokatlan: „A hős, a gonosz és a szent” (Hero, villain and saint). Ebben az esszé-kötetben Ön szinte irodalmi eszközökkel boncolgatja az emberi fejlődés filozófiai és pszichológiai problémáit. Hogyan tudja Ön összeegyeztetni alkotó munkásságában a „két kultúra” megközelítésmódjait?

— Ahogy mondani szokták: ez jó kérdés. Pályafutásomat fizikusként kezdtem. Azért változtattam pályát, mert úgy éreztem, hogy a kutatólaboratórium távol tart az igazi élettől. A pszichoanalízis felé fordultam és problémás gyerekek kezelésével foglalkoztam, miután annak rendje-módja szerint kiképzést nyertem a gyermekpszichoanalízisben. Nyolc évig dolgoztam Bruno Bettelheim mellett, aki, mint bizonyára tudja, Ausztriából került az Egyesült Államokba. Sok skizofrén gyerekekkel foglalkoztam abban a nyolc évben. Nagyon érdekes munka volt ez, és rengeteget tudtam meg közben önmagamról és az élet gazdagságáról. Rá kellett azonban jönnöm, hogy nem vagyok igazán jó therapeuta. Túlságosan élénk, túl energikus voltam pácienseim számára, akik nem tudtak mit kezdeni energiáimmal, én pedig nem tudtam elleplezni a bennem feszülő energiát. A matematika még mindig nagyon érdekelt, fel is használtam a vizsgálati adatok elemzésére, és különböző kutatócsoportok munkájában is közreműködtem matematikai statisztikai tanácsadóként. Így jutottam a Chicagói Egyetemre, a statisztika tanárának. És bár a fejlődéslélektan keretében továbbra is tartok előadásokat a pszichoanalízis témaköréből, a méréselmélet egyre jobban lekötött, és a legutolsó egy-két évtizedben leginkább ez foglalkoztatt.

A Hős-könyv pszichoanalízis tárgyú előadásaimból keletkezett, és tulajdonképpen tizenöt évvel ezelőtt írtam, de csak most publikáltam — majdnem véletlenszerűen, a kiadó unszolására. Örülök, hogy most elővet-

tem, mert rájöttem, hogy még mindig nagyon érdekel a téma, noha ma már nem dolgozom ez irányban. Amit sajnálok is egy kicsit.

A pszichoanalízissel egyébként gyermekkoromban találkoztam először, amikor hét éves kisfiúként egy esztendeig engem is analízisbe járat-tak, mivel nagyon kiborított családunk helyzete, és mivel nehezen tanul-tam olvasni. Nem is tanultam meg nyolc vagy kilenc éves korom előtt, és mindig lassú és nehézkes olvasó maradtam, bár olvasni szeretek. Közép-iskolásként nagyon jól ment viszont a matematika és a fizika, és ez vitt a fizikus pályára. A Bell Laboratóriumban a későbbi Nobel-díjas Charles Townes munkatársa lettem, aki azt szerette volna, ha követem a Colum-bia Egyetemre, és ott szerzem meg a Ph. D.-met. De én inkább Chicagót választottam, ahol egy másik leendő Nobel-díjas, Robert Mulliken mellett dolgoztam, aki a kémiai kötések energiáit és a molekulák elektronszerke-zetét kutatta. Miként azonban már mondtam, nem tudtam elviselni a la-boratóriumi munka elszigeteltségét — a szociális elszigeteltséget —, a va-ló életbe akartam belemerülni, és így kötöttem ki a pszichoanalízisnél és a pszichikus zavarokkal küszködő gyermekeknél. És azt hiszem, tulajdon-képpen azóta próbálkozom azzal is, hogy kikutassam, miként lehet mérni a gazdagon tartalmasat, azt, ami jelentéssel telített. Sajnos mégsem halad-tam ebben elég messzire, azt hiszem.

— Milyen távlatai vannak Ön szerint a Rasch és mások által kezde-ményezett új módszertani törekvéseknek?

— Azt hiszem, a társadalomtudományok fontos fordulat előtt állnak, és a Rasch által megalapozott új módszerek belátható időn belül a min-dennapi munka megszokott eszközeivé lesznek. Úgy gondolom, hogy a társadalomtudósok következő nemzedéke magától értetődően sajátítja majd el ezeket a módszereket, és ugyanolyan természetességgel fogja fel-használni őket adatai elemzésére, mint ahogyan most az elemi statisztikai eljárásokat használja. Úgy gondolom továbbá, hogy a kutatók a Rasch-féle mérési módszereket hamarosan adottnak tekintik, és olyan magától érte-tődően alkalmazzák majd, mintha a társadalomtudományok mindig is ezt tették volna. A fordulatnak pedig az lesz a következménye, hogy végre a társadalomtudományokban is meghonosodnak a jól használható, szabatos, reprodukálható és általánosan elfogadott változók, amelyek a pszicholó-giai, pedagógiai, szociológiai szituációkat leíró, tartós és érvényes tudás szilárd építőkövei lesznek.

C. LEWIS: A GONDOLATOK CSERÉJE ERŐSÖDIK

— Hogyan válik valaki Amerikában a pszichometria szakértőjévé? Ön miként kötött ki a pszichometriánál?

— Azt hiszem, különféle véletlenek következtében. Még a pszichometria szót sem hallottam, mielőtt a felső évfolyamokra, a graduate schoolba kerültem, de mint alsós, azaz undergraduate matematikára szakosodtam, és ezenkívül a pszichológia egyes ágai, különösen az érzékelés és az alaklélektan iránt érdeklődtem. De szinte egyáltalán nem tanultam statisztikát, tesztelméletet vagy mérés- és skálaelméletet. Felsőbb évfolyamokra jutva mégis (az ETS által is szponzorált) pszichometriai szakirányra jelentkeztem a Princetoni Egyetemen, és 1965-ben kezdtem el tanulmányaimat az egyetem pszichológiai intézetében. Egy bizonyos ponton aztán felismertem, hogy a statisztikai vonatkozások jobban érdekelnek, mint a pszichológiaiak. Éppen ekkoriban hozott létre az egyetemen egy új statisztikai intézetet John Tukey, és így átiratkoztam oda, az ETS-szel való kapcsolatomat is fenntartva. Ez aztán egészen jól bevált nálam. Azóta is megőriztem kötődésem a pszichológiához, sokáig dolgoztam pszichológiai tanszékeken; de fő kutatási területem a statisztika pszichológiai alkalmazása maradt, persze nem csupán a tesztelméletre korlátozva, hanem a kísérlettervezés és varianciaanalízis, a lineáris modellek problémái is foglalkoztattak. Az egyik fő érdeklődési területem azonban kétségkívül a pszichometria volt, és természetesen az ma is.

— Ön olyan kiváló tudósoknak volt tanítványa és munkatársa, mint Fred Lord és Melvin Novick. Mi volt az ő szerepük a tesztelmélet továbbfejlesztésében?

— Talán el kell még mondjam azt, hogy ha már egyetemistaként elkezdtem pszichometriával foglalkozni, abban igen nagy szerepe volt annak, hogy olyan kiváló emberektől tanulhattam, mint Fred Lord és Mel Novick, továbbá Harold Gulliksen, aki az ETS-ben irányította a pszichometriai programot, valamint Karl Joreskog, aki akkoriban meghívott professzor volt a Princetoni Egyetemen és az ETS-ben is.

Rátérve Fred Lordnak és M. Novicknek a pszichometria fejlődésében játszott szerepére, az az érzésem, hogy a tesztelmétről 1968-ban megjelent könyvük hozta ezt a területet érett fázisba — azt hiszem, ez a legmegfelelőbb szó. Gulliksennek a hatvanas évek elején publikált munkája ugyancsak rendkívül fontos lépés volt ebben a fejlődésben, mert különféle megközelítési módokat összefogott, és összefoglalta azokat az eredményeket és tapasztalatokat, amelyeket az előző évtizedek felhalmoztak. De csak Lord és Novick könyve volt az a munka, amely a különböző

megközelítésmódoknak szisztematikus és szigorú matematikai statisztikai megalapozást adott, és tisztázta, szabatosan meghatározta az olyan fogalmak tartalmát is, mint a valódi pontérték és a mérési hiba. Persze, ami a legemlékezetesebbé teszi ezt a könyvet, az azoknak az elsősorban Allan Birnbaumtól — de persze Fred Lordtól is — eredő elgondolásoknak a bemutatása, amelyek később Item Response Theory (IRT) néven terjedtek el. Elméleti szempontból ez ma a domináló irányzatok egyike, és ez a domináns szerep nagyon szépen visszavezethető LORD és NOVICK 1968-ban publikált könyvének hatására. Amivel nem azt akarom mondani, hogy nem voltak e területen korábbi publikációk is — ide sorolhatnám F. Lord előző monográfiáit, Rasch könyvét és A. Birnbaum kutatási jelentéseit —, de mindezeknek a kezdeményezéseknek egyetlen szisztematikus keretben való egyesítése és a tesztmérések szélesebb összefüggésekbe helyezése a Lord—Novick könyv teljesítménye volt. Lord és Novick munkássága persze nem korlátozódott erre az alapvető fontosságú műre. Lord igen nagyszámú és kiváló tanulmányokat publikált e területen azóta is, és igen változatos összefüggésekben nagy hatást gyakorolt a pszichometriára mind az ún. klasszikus tesztelmélet, mind pedig az item—válasz elmélet (IRT) vonatkozásában. Lord egyik legfontosabb eredménye kétségtelenül az adaptív (vagy, ahogy korábban nevezték: egyénre szabott, tailored) tesztelés elgondolása, amely szerint a tesztelés folyamatának egy bizonyos pontjától kezdődően az egyén által megválaszolendő itemet az előző itemekre adott válaszáról függően választják ki. Vagyis a tesztelés a megválaszolendő itemek kiválasztását tekintve adaptív, az egyén szintjéhez igazodó. Ez az eljárás a gyakorlatban csak a számítógép felhasználásával — számítógéppel „adagolva” — valósítható meg, de Lord már akkor felvázolta elméletileg, amikor a komputerrel adagolt tesztek még a realitástól távol eső álomnak tűntek. Manapság viszont az ETS-nek is, más tesztszervezeteknek is vannak adaptív tesztjei, és a tervek szerint ezt a tesztformát a jövőben jelentősen kiterjesztik, felhasználva őket olyan nagyszabású programokban is, mint (az európai egyetemi államvizsgálóhoz hasonlítható) GRE-vizsgák (Graduate Record Examinations).

Lordhoz hasonlóan Novick is számos terület fejlődésére hatott azon kívül is, hogy szerzőtársa volt Lordnak. Különösen vonatkozik ez arra a törekvésére, hogy a bayesiánus szemléletmódot alkalmazza a tesztelméleti problematikára. Olyan program volt ez, amelyen rendkívül intenzíven és eredményesen dolgozott a hatvanas évek közepétől egészen az 1986-ban bekövetkezett haláláig. Egyebek mellett Novick vezette be a pszichometria világába (Lindleyvel, Jacksonnal és másokkal együttműködve) azokat a módszereket, amelyeket ma „empirikus alapú becslési eljárásoknak”

(empirical based methods of estimation) neveznek, és amelyekben más csoportoktól vagy tanulóktól nyert információkat hasznosítanak egy újabb csoport vagy tanuló értékelésekor. Bayesiánus alapokon a reliabilitás becslésére is új megközelítésmódot talált, és a bayesiánus hasznossági (utility) elméletre alapozva a torzításmentes tesztelés elméletén is dolgozott (Nancy Petersonnal közösen), és ezek az elgondolásai is nagy hatással voltak a későbbi fejlődésére.

Úgyhogy elmondhatom, mind Lord, mind pedig Novick igen jelentős szerepet játszott a pszichometria újabb haladásában.

— Ön jónéhány évig egy európai egyetemen, Hollandiában működött. Személyes tapasztalatai szerint vannak-e speciális, az amerikaiaktól eltérő vonásai a tesztelméletnek Európában — és mennyire ismerik, illetve befolyásolják egymást az európai, illetve amerikai kutatási irányzatok?

-- Legelőször is azt mondhatom, hogy csakugyan rendkívül erősen befolyásolják egymást, annyira, hogy nehéz volna valóban elkülönülő, egymástól független fejlődésről beszélni. Az irányzatok egymásra találásának egyik korai lehetőségét éppen az a Fred Lord által életre hívott program biztosította, amely révén fiatal európai tudósok egy-két évi kutatómunkára az ETS-be jöhettek, és ott maguk is befolyásolták a kutatómunkát. Hosszú listájukon olyan nevek találhatók, mint Karl Joreskog, Michel Brown, Walter Kristoff, Erling Andersen — és a névsor még folytatódik. Vagyis az ETS, és személy szerint F. Lord is hozzájárult a két fejlődési vonal egymásra hatásához.

Amikor azonban Hollandiában megkezdtem egy évre tervezett és kilenc évre sikeredett munkámat, szembetűnő volt számomra bizonyos szemléletbeli különbség is, ez pedig főként az európai kollégáknak az elméleti megalapozottság és matematikai szigor iránti erősebb hajlamukban és igényükben mutatkozott meg. A legjobb példa erre talán az, hogy az IRT alkalmazása a Rasch-modellhez kapcsolódó munkákra korlátozódott. A dán statisztikus által kidolgozott tesztelméleti módszertan kétségtelenül Hollandiában, Németországban, Ausztriában és bizonyos mértékig Angliában talált a legtöbb követőre. E módszer átfogó filozófiai-tudományelméleti megfontolásokra támaszkodik — Rasch és mások a mérések „specifikus objektivitását” tűzték ki célul —, és a hollandiai (és más európai) kollégák közül sokan azt állítják, hogy csakis a Rasch-modellre épülő méréselmélet lehet az egyedül értelmes és eredményes, az összes többi nem koherens, nem valóságos méréseket szolgáltat. Ez a felfogás az én szememben túlon túl extrém, de ismétlem, nagyon sok hollandiai és

más európai országbeli kollégának nézete ez. Természetesen a Rasch-moddellnek megvannak a maga vonzó tulajdonságai, mint például matematikai egyszerűsége, statisztikai kezelhetősége, és e tulajdonságokat jól használták ki az e területen dolgozó kutatók.

Alapvető beállítódásuknak megfelelően az európai kutatók a modell és az adatok illeszkedése kérdéseinek is sokkal több figyelmet szentelnek annál, mint amit én az USA-ban láttam. Ide tartozik a Rasch-moddell illeszkedésének problematikája is, és azok a módszerek, amelyekkel a nem illeszkedő itemeket kiszűrik minden olyan itemhalmazból, amellyel egy konkrét skálát vagy látens dimenziót próbálnak meghatározni. Az Egyesült Államokban tapasztalataim szerint a kutatók sokkal pragmatikusabb beállítódásúak, ami először is abban nyilvánul meg, hogy a modellek sokkal szélesebb skáláját tartják elfogadhatónak: a Rasch-moddelt is alkalmazzák bizonyos körülmények között, más összetettebb modelleket használnak fel más körülmények közt, és inkább hajlandók elfogadni a modellhez való illeszkedés bizonyos fogyatékoságait. Figyelmüket arra összpontosítják, hogy használható tesztek nyerjenek, olyat, amely megfelelőképpen jellemez egy adott szituációt akkor is, ha nélküli a magasfokú elméleti tisztaságot.

Egy újabb keletű terület, ahol olyan fejleményeket tapasztalok, hogy egymástól viszonylag függetlenül keletkezett hollandiai és amerikai kutatási irányok hatnak egymásra, az az ún. automatizált teszt szerkesztés. Holland kutatók e téren született kezdeményezéseire számos európai és amerikai szakember (köztük kollégám, M. Stocking) figyelt fel és folytatta, módosította, kiterjesztette elgondolásaikat.

Az amerikaiak számára általában is hozzáférhetőbbé tette az európaiak munkásságát az utóbbi éveknek az a gyakorlati fejleménye, hogy a legtöbb európai szerző egyre inkább törekszik újabb munkáinak angol (gyakran amerikai angol) nyelvű megjelentetésére; másrészt az európaiak változatlanul élénk érdeklődéssel kísérik amerikai kollégáik működését, úgy, hogy a gondolatok cseréje mindkét irányban fokozódik.

— Az ön kutatócsoportja milyen feladatokon dolgozik jelenleg? Mik a várható újabb fejlemények e tudományterületen?

-- Talán azzal kezdeném, hogy minden munkatársamnak legalább kétféle típusú kutatási tevékenysége van. Az egyik, amelyiket alapkutatásnak nevezhetünk, többé-kevésbé saját, egyéni kezdeményezésű, a kutató problémaérzékelésétől és személyes érdeklődésétől függ; a másik pedig

inkább alkalmazott jellegű és gyakorta az ETS-n belül konkrét programok vagy csoportok igényeit elégíti ki, szoros együttműködésben dolgozva valamilyen fejlesztési feladaton. Azt mondtam, hogy ez két különböző kategória, de igazából gyakori az átfedés, mert gyakran előfordul, hogy az alkalmazói igények egybeesnek az alap kutatások irányával.

Hosszú lenne részleteznem a témákat, amelyekkel munkatársaim jelenleg foglalkoznak. Ehelyett csak néhány jellegzetes példát említenék.

Bob Mislevy, akinek a nevéhez fűződik a NAEP (National Assessment of Educational Progress), vagyis az amerikai közoktatás eredményességét ellenőrző tesztprogram számos eszközének a kidolgozása, újabban egy határterületre összpontosítja figyelmét: azt vizsgálja, hogy miként tükröztethetők a pszichometriában a kognitív pszichológia eredményei. E kutatásaiban együttműködik az ETS más részlegeiben tevékenykedő kutatókkal, akiknek szakterülete közvetlenebbül kapcsolódik a kognitív tudományokhoz. Arra törekszenek, hogy az IRT egyszerű fogalmainak, a változók mögött húzódó látens tulajdonságok fogalmának helyébe egy sokkal összetettebb, kidolgozottabb modellt léptessenek, amely figyelembe veszi a különböző képességeknek és ismereteknek a válaszolást meghatározó hatását is. Ehhez felhasználják azokat a statisztikai eljárásokat, amelyeket az ún. döntési fák elemzésére dolgoztak ki.

Ugyancsak a hagyományos pszichometria és a kognitív tudomány összekapcsolásán fáradozik Kikumi Tatsuoka, aki még Chicagóban kezdte ilyen irányú kutatásait. Egyebek mellett az elkövetett hibák, a hibás választok elemzésére dolgozott ki módszereket. Tanulmányozza a modellhez nem illeszkedő, abból „kilógó” egyének teljesítményének jellegzetességeit, mivel olyan specifikus kognitív folyamatok megnyilvánulásainak tartja a modellhez nem illeszkedő szabálytalanságokat, amelyek specifikus ismeretstruktúrákra épülnek. Kikumi Tatsuoka kutatásai összekapcsolják az értékelési folyamatokat az oktatás—tanulás folyamatával. Olyan törekvés ez, amely újabban az egész ETS-t is foglalkoztatja.

Megnyilvánul ez a törekvés Kentaro Yamamoto kutatásaiban is. Ő mostanában a felnőttek írni-olvasni tudásának mérésébe kapcsolódott be, ahol újszerű itemtípusokkal vizsgálják az írás-olvasás különböző aspektusait, és ezeket a nem hagyományos itemtípusokat kombinálják IRT-modellekkel, hogy mélyebben elemezhesék az írás- és olvasási nehézségek fajtáit. Kentaro dolgozta ki az ún. hibrid modellt, amely a látens osztályok elméletével kapcsolja össze a szokványosabb IRT-modelleket. De itt említhetném meg Peter Pashley alternatív három paraméteres modelljét is.

Néhány további fontos és jellegzetes kutatási terület, csak címszavakban: a tesztek összemérhetőségének (equating) feltételei (G. Boodo, H. Brown, K. Sheehan), a pedagógusok szakmai alkalmasságának újszerű vizsgálata (M. Stocking, G. Boodo stb.), komputerizált tesztelés (M. Stocking, K. Sheehan, Duan Li stb.), a teszt-dimenziók vizsgálata az IRT eszközeivel (Ming May Wan) stb.

A látszólag külön területeken mozgó kutatások is gyakran összekapcsolódnak módszertani vonatkozásban, például a bayesiánus szemlélet és eszközrendszer alkalmazásával.

— Mondana valamit a saját kutatásairól is?

— Mostanában a komputerizált szakmai alkalmassági (mastery) teszteléshez kapcsolódó konkrét feladaton dolgoztam K. Sheehannal és másokkal együtt. Építészeti szakvizsgáiról volt szó, amelynek teljesen komputerizált formát adtunk. A számítógép teljesen új feladat típusok alkalmazását tette lehetővé, ahol is a feladat tényleges építész-mérnöki tevékenységet szimulálhat a CAD (computer aided design; komputerizált műszaki tervezés) felhasználásával. De a számítógép más lehetőséget is kínált: a feladatok nehézségi szintjét az egyénhez igazítottuk. Amihez persze megfelelő matematikai módszereket is kellett találni, e módszereket azután más összefüggésekben is felhasználhattuk.

Ez a téma számomra a külső megrendelésre készülő, alkalmazott kutatás jellegű feladatnak és alapkutatási érdeklődésemnek nagyon szerencsés találkozását hozta. Matematikai statisztikai szempontból ez az építész-tesztelés ugyanis hipotézisek szekvenciális tesztelését, bayesiánus módszerek felhasználást jelentette.

Ugyancsak mostanában visszatértem egy másik témához, amely — igaz, más, eredetileg pszichofizikai összefüggésben — már disszertációm megírása óta elkísér. Jelenlegi formájában a nem paraméteres IRT-modellekről van szó. Hollandiában abban a szerencsében volt részem, hogy együtt dolgozhattam Robert Mokkennel, akit szakmai körökben egy nem paraméteres IRT-skála kifejlesztőjeként ismernek, s aki a nem paraméteres modellilleszkedés vizsgálatára is módszereket adott. Közös munkánkban annak idején engem főként az egyénre vonatkozó következtetések bizonytalanságainak becslése foglalkoztatott, ami bayesiánus szemszögből egy a posteriori eloszlás meghatározását jelentette, méghozzá az adott összefüggésben olyan eloszlás megadására volt szükség, amely intervallumokkal operált, látens osztályokra vonatkozott. Újabb munkámban az ite-

mek nehézségi sorrendjének bizonytalanságaira is kiterjesztettem a vizsgálódásaimat. Ez így bizonyára nagyon elvontan hangzik — bizonyos értelemben csakugyan az —, de összefügg az előbb említett konkrét feladattal is.

— Engedjen meg még egy utolsó kérdést: jelenleg milyen általános fejlesztési törekvések jellemzik az Educational Testing Service tevékenységét?

— Az ETS olyan korszakában van, amelyet a megújulásra való törekvés jellemez. Több irányú törekvés ez, amelynek jónéhány példáját már érintettem saját kutatásainkat ismertetve. Ez persze nem véletlen egybeesés...

Említettem például a komputerizált vizsgáztatással való kísérletezést, és azt a körülményt, hogy a komputer nemcsak a feladatok új szisztémájú adagolását teszi lehetővé, hanem elvileg új típusú feladatok megjelenését is. Bár az ETS továbbra sem veti el a feleletválasztásos itemeket, a legmegszokottabb feladattípust, mely a maga helyén jól használható, mégis általános az a törekvés, hogy eszköztárát más, alkotóbb feladatokkal is kiegészítse.

Az ilyen irányú kísérletezésnek és újításoknak különösen fontos terpei a szakmai működési bizonyítványok tesztvizsga-rendszerei. Ide sorolhatók a már említett építész- és pedagógus-tesztvizsgák is.

A pedagógus-tesztrendszer kifejlesztése egyike az ETS-ben igen ritka belső programoknak. Vagyis olyan kutatási programról van szó, amelyet külső megrendelés és külső anyagi támogatás nélkül kezdtük el. Az ETS azonban úgy érzi, hogy e tesztrendszer kifejlesztésével jótékonyan befolyásolhatja a közoktatást. Újszerű ez a tesztrendszer annyiban is, hogy nemcsak afféle záró követelményrendszert tartalmaz, hanem az alkalmaság előzetes megállapításától kezdve a pedagógusképzés különböző szakaszait is felöleli, segíti a képzés egész folyamatát.

A közelmúltban fejeződött be az egyik legismertebb, legszélesebb körben alkalmazott tesztrendszerünk, az érettségit helyettesítő SAT (Scholastic Aptitude Test) több évig tartó felülvizsgálata, modernizálása is, ahol szintén megszüntettük a feleletválasztásos itemek korábbi kizárólagosságát.

Általánosságban elmondható, hogy az ETS a legutolsó négy-öt évben nem éri be a meglévő (és többnyire jól bevált) tesztprogramok karbantar-

tásával, csiszolgatásával, hanem olyan új tesztprogramok kifejlesztéséhez látott hozzá, amelyek túlmutatnak az ETS hagyományos érdeklődési és tevékenységi körén, és az oktatás folyamatának közvetlenebb befolyásolásával is szolgálják a közoktatás színvonalának emelését. Az értékelési és a tanulási folyamatok integrálását igényli ez a szemléletváltás, amely a kutatási feladatok szempontjából a kognitív pszichológia jelentőségének felértékelődésével járt együtt.

G. H. FISCHER: MEGNŐTT AZ EGZAKTSÁG IRÁNTI IGÉNY

— Ha nem is egyedülálló, mégsem túlzottan megszokott jelenség, ahogyan a bécsi egyetemen a matematikai pszichológiának egy jellegzetes iránya igen erősen van képviselve. Hogyan alakult ki ez a helyzet, milyen hagyományok vagy elődök járultak hozzá ehhez a fejlődéshez?

— Amikor 1966-ban tanársegédként visszakérültem a bécsi egyetem pszichológiai intézetébe, ott különösebb tradíciója nem volt a szó szorosabb értelmében vett matematikai pszichológiának vagy a pszichometriának. Ezt megelőzőleg néhány évet az IHS-ben tevékenykedtem. (IHS: Institut für Höhere Studien und Wissenschaftliche Forschung. Ezt a hasonló nevű princetoni intézet mintájára létesült kutató és továbbképzési intézményt a Ford Alapítvány és Bécs városa közösen finanszírozta.) Az IHS-ben, amely teljesen társadalomtudományi irányú volt, egyebek közt attitűdméréssel foglalkoztam, és a híres Paul Lazarsfeld professzorral is megismerkedtem, aki akkoriban vendégtanárként jött Bécsbe. Az ő hatására figyeltem fel a „látensstruktúra-analízisre”. Ezek a tapasztalatok bizonyára befolyásolták későbbi tudományos fejlődésemet. E ponton talán meg kell említenem, hogy a húszas években Lazarsfeld is a bécsi egyetem pszichológiai intézetében kezdte el tudományos pályafutását, és pedig Karl és Charlotte Bühler vezetésével. A látens struktúra analízisét és a látens osztályok analízisét azonban jóval később, az USA-ban fejlesztette ki. Ez magyarázza, miért mondom, hogy a bécsi pszichológiai intézetben ilyen jellegű tradíció nem élt elevenen.

— Ön személy szerint inkább pszichológusnak vagy inkább matematikusnak tartja magát? Hogyan sikerült tanulmányai és munkássága során ezt a két különféle — humán és természettudományos -- érdeklődési irányt összeötvöznie?

— Nem szeretném magamat sem a „pszichológus”, sem a „matematikus” címkével felcímkézni; tudományos — tevékenységemre leginkább a „pszichometrikus” megjelölés illik rá, abban az értelemben, ahogyan ezt

a szót az USA-ban általában használják. Ami az egyetemi tanulmányaimat illeti: eredetileg a pszichológia volt az első szakom, és fizika lett volna a második. A fizikához azonban magasabb matematikai előismeretek is kellett; ezért aztán elkezdtem járni a matematika-előadásokra is, és otthagadtam.

A Rasch-modellel miként ismerkedett meg, és mi vezette el e modell jelentőségének korai felismeréséhez?

— A rábukkanás a Rasch-modellre véletlennek volt köszönhető: eredetileg a faktoranalízisről akartam könyvet írni, és egyet s mást publikáltam is ebből a tárgykörből J. Ropperttel közösen. (1963-ban ortogonális és ferdeszögű Prokrusztész-eljárást alakítottunk ki, ami a szakirodalomban akkoriban újdonság volt, és Monte-Carlo szimulációs módszereket is alkalmaztunk már a faktoranalízis vizsgálatára.) Minthogy azonban 1966-ban Kurt Pawlik megjelentette monográfiáját a faktoranalízisről („Dimensionen des Verhaltens”), és hát a német nyelvterület tudományos „piaca” akkoriban igen korlátozott volt, a pszichológia egy másik, kellőképpen formalizált területe után kutattam, és rá is leltem arra, amit kerestem, a tesztelméletben. Az akkoriban domináló „klasszikus tesztelmélet” azonban száraznak és terméketlennek tűnt számomra. Véletlenül szereztem tudomást egy matematikai pszichológiai szemináriumról, amelyet egy holland alapítvány, a „Netherlands Universities Foundation for International Cooperation” (NUFFIC) rendezett Hágában, 1966 júliusában. A programban neves professzorok, R. J. Audley, W. S. Torgerson, C. H. Coombs, L. J. Cronbach előadásain kívül egy számomra teljesen ismeretlen Dr. G. Rasch neve is szerepelt. Jelentkezésemet elfogadták. A szeminárium munkarendje olyan volt, hogy szinte az egész napot kitöltötték az előadások, s emellett még irodalmat is kaptunk tanulmányozásra. Ez túlságosan sok volt, ezért aztán hamarosan kialakult, hogy ki-ki csak néhány előadást látogatott. Az ismeretlen G. Rascht kevesen választották, meg azután nem is volt könnyen érthető, amit mondott; így aztán előfordult, hogy a foglalkozásain egyedül voltam jelen. Bár a matematikai részleteket nem mindig tudtam követni, Rasch és modellje mégis fellelkesített. Úgy éreztem, hogy a pszichológiai mérés problémáinak ez a formalizálása mélyebb és adekvátabb, mint a korábban meghonosodott mérési és skálaszerkesztési módszerek. Egyik beszélgetésünk alkalmával Rasch javasolta, hogy menjek el néhány hónapra Koppenhágába. A következő évben azután két hónapot ott töltöttem Raschnál, a Statistike Institutban. Ott P. Alleruppal közösen megírtuk első változatban azt a FORTRAN-programot, amely a Rasch-modell itemparamétereit feltételes maximum likelihood

(CML) becslési eljárással számította ki, és amelyet Bécsbe való visszatérésem után még lényegesen tovább javítottam. 1968-ban az általam szerkesztett „Psychologische Testtheorie” egyik fejezeteként Alleruppal közösen publikáltuk is ezt a programot, amely lényegesen hozzájárult a Rasch-modell elterjedéséhez a német nyelvterületen. Már akkoriban is érdekelt az a probléma, hogy a Rasch-modell levezethető-e és miként vezethető le általánosabb előfeltételekből; az 1968-as könyv egy másik fejezetében bizonyítást adtam meg arra, hogy a Rasch-modell a nyerspontértékek elégséges statisztika voltából (és még néhány, az itemjelleg-görbékre vonatkozó technikai feltételből) levezethető.

— Melyek voltak az első lépések e modell alkalmazásában és továbbfejlesztésében?

— 1968-ban habilitáltam statisztikából a bécsi egyetemen, és még abban az évben kineveztek a pszichológiai intézet pszichológiai módszertani és matematikai pszichológiai tanszékére professzornak. 1969-től két tanársegédem is volt, H. Spada és H. Scheiblechner. Együttműködésünk termékeny tudományos klímát teremtett, amelyben sokféle alkalmazás és továbbfejlesztés gondolata vetődött fel. Utólag visszatekintve — a dichotom és polytom Rasch-modellhez kifejlesztett programjainkon kívül — különösen azok a kezdemények tűnnek fel a legmegtermékenyebbítőknak, amelyek az itemparamétert további, még elemibb „bázisparaméterekre” bontották fel (LLTM, LLRA). Ezek a modellek az itemtartalmak „facet”-elemzésében és terápiás hatások mérésében is rendkívül hasznosnak bizonyultak.

— Hogyan fejlődött ez a tevékenység a továbbiakban?

— Röviden aligha tudnám összefoglalni a későbbi fejleményeket általában, ezért inkább arra szorítkoznék, hogy miként alakult tovább a mi munkánk itt Bécsben. Egyfelől sok és sokféle alkalmazásra (és publikációra) került sor a tesztanalízisben, a tananyagelemzésben és a változásdiagnosztikában. Csak néhány nevet említenék egykori munkatársaim és tanítványaim közül: A. Formann, K. Kubinger, W. Kempf, I. Rop; de még sorolhatnám. Másrészt tovább javítottuk a becslési algoritmusokat és a computerprogramokat. Némelyik ezek közül az én 1974-es könyvemben („Einführung in die Theorie psychologischer Tests”) jelent meg, mások a Kempf és Repp által szerkesztett tanulmánygyűjteményben („Mathematical Models for Social Psychology”, 1977) és a Spada és Kempf által szerkesztett kötetben („Structural Models of Thinking and Learning”, 1977) vált széles körben is hozzáférhetővé. Az én számomra pedig egyre fonto-

sabbá vált az a kérdés, hogy miként alapozhatók meg matematikailag a Rasch-modellek és kiterjesztéseik: Milyen feltevésekből vagy „axiómákból” következnek ezek a modellek matematikai szükségsszerűséggel? Ha sikerül ilyen axiómarendszert találni, nem kell a modell további matematikai megalapozására törekedni vagy alternatívák után kutatni.

Törekvésemben az a megfigyelés is vezetett, hogy minden olyan újítás, amit nem sikerült matematikailag teljesen egzakt módon megalapozni, rövidesen el is tűnt a szakirodalomból. Csak azok a módszerek és eredmények álltak ki az idők próbáját, amelyeket matematikailag is igazolni lehetett. Azonkívül egyre több hivatásos statisztikus és matematikus dolgozik és publikál az Item Response Theory területén, és ezek a kollégák csak azt fogadják el, ami matematikailag egzaktul bizonyítva van. Más szóval jelentősen megnőtt az egzaktság iránti igény. Ebből kiindulva a legutóbbi évtizedben különösen sokat foglalkoztattak az egyértelműségi és identifikációs problémák, valamint a Rasch-modellek skálatulajdonságai. E kérdések némelyikét a Psychometrika hasábjain (1981, 1983, 1989) megválaszoltam.

Máshelyütt a pszichometrikai irodalomban gyakran félreértett fogalmat, a „specifikus objektivitást” és következményeit, különösen a Rasch-modell skálatulajdonságait elemeztem (Psychometrika 1987, és K. KUBINGER szerk. „Moderne Testtheorie”, 1988).

— Milyen szerepet játszott a bécsi egyetem a modern tesztelmélet európai elterjesztésében?

— Ezt mások jobban meg tudják ítélni, mint én. Nyilvánvaló, hogy két könyvem („Psychologische Testtheorie”, 1968 és „Einführung in die Theorie psychologischer Tests”, 1974), továbbá az a körülmény, hogy H. Spada 1972-ben Bécsből Kielbe, az Institut für die Pädagogik der Naturwissenschaftenbe került és ott széles körben alkalmazta pedagógiai vonatkozásban a Rasch-modelleket, legalábbis ismertebbé tette ezt a problémakört.

— Mi a véleménye az ún. Item Response Theoryról és a tesztfejlesztés amerikai gyakorlatáról?

— Úgy gondolom, hogy az USA-beli és az európai fejlődés között valóban volt bizonyos eltérés, legalábbis ami a múltat illeti. Ez szerintem a következőkből adódott: Európában általában és a német nyelvterületen különösen az Item Response Theoryt elsősorban a pszichológiai tesztelés

és a pszichológiai mérések megalapozására szolgáló eszköznek tekintették. A Rasch-modell ezért olyan ideálnak számított, amelyhez a pszichológiai tesztek a lehetőség szerint közelíteni kell. USA-ban a pedagógiai célú tesztelés („Educational Testing”) korábban és erőteljesebben kifejlődött, mint Európában, a tesztek nagyobb szerepet kaptak az iskolai értékelésben és az egyetemi felvételen. Az itemeket — tisztán praktikus okokból — szinte mindig feleletválasztásos formában adták meg. A pszichometrikusoknak inkább az adatelemzői szerepkör és nem a teszt szerkesztői feladat jutott. Ezért inkább arra törekedtek, hogy a tesztmodelleket igazítsák az adott adatokhoz és ne fordítva, a tesztek egy meghatározott modellhez. Ez lehet az oka a két és három itemparaméteres modellek USA-beli népszerűségének. Egy kollégánk kihegyezett megfogalmazása szerint az amerikaiak inkább „pragmatikusok”, az európaiak pedig inkább „fundamentalisták”. Az utóbbi időben azonban az álláspontok közeledése figyelhető meg.

— Melyek a jelenlegi kutatási céljai? Legújabb eredményei?

— Algoritmusokat és szoftvert fejlesztünk ki a polytom modellek egy osztályához (Linear Rating Scale Model, LRS; lásd FISCHER és PARZER, *Psychometrika*, 1991; Linear Partial Credit Model, LPCM, előkészületben), amelyek a polytom Itemek vagy „rating”-skálák tanulmányozására alkalmasak, másrészt a dichotom Rasch-modellt, valamint az LLTM-t és LLRA-t is mint speciális eseteket felölelik. Vagyis sok különféle alkalmazási területhez ugyanaz az egy-két program elegendő lesz. Ezek a fejlesztések egyúttal azt is lehetővé teszik, hogy a modellek a korábbinál hosszabb tesztekre és ezzel újabb alkalmazási területekre is kiterjeszthetők legyenek. Elméleti szempontból változatlanul a modellek axiómarendszerekből való levezethetősége érdek. Hasonló kérdéseket a hatványsor-modellekkel és Poisson-modellekkel kapcsolatban is vizsgáltam (lásd Doignon és Falmagne szerk. „Mathematical Psychology: Current Developments”, 1991).

— Milyen további fejlődést vár a tesztelmélettől?

— Nem szívesen vállalkoznék jóslatokra a további fejlődés irányait illetően; egy-egy új szempont felmerülése meglepően új irányokba terelheti a kutatást.

— Milyen jelentősége van a modern tesztelméleti módszereknek általában a pszichológia elmélete és gyakorlata számára?

— Általában a pszichológiára tett hatását nem szeretném eltúlozni, már csak azért sem, mert a pszichológusok derékhada egyre nehezebben tudja követni az Item Response Theory mind bonyolultabbá váló fejlődését. Egy kissé tartok is tőle, hogy ezt a területet a jövőben egyre inkább a statisztikusok és a matematikusok veszik át, és ennek következtében a pszichológiából kiválik. A gyakorlati tesztp Praxisban az item Response Theory oly mértékben tesz majd szert egyre nagyobb jelentőségre, amilyen mértékben elterjed a computerizált tesztelés.

A kézirat elfogadva: 1992. március

IRODALOM

- ANDERSEN, E. B., 1970, Asymptotic properties of conditional maximum likelihood estimators, *J. of the Royal Statistical Society*, 283—301.
- ANDERSEN, E. B., 1971, The asymptotic distributions of conditional likelihood ratio tests, *J. of the American Statistical Association*, 630—633.
- ANDERSEN, E. B., 1983, Analyzing data using the Rasch Model, In: ANDERSON, S. B., HELMICK, J. S. (eds.), 1983, *On educational testing*, San Francisco, Jossey-Bass, 193—223.
- ANDERSON, S. B., BALL, S., MURPHY, R. T., 1975, *Encyclopedia of educational evaluation*, San Francisco, Jossey-Bass.
- ANDERSON, S. B., HELMICK, J. S. (eds.), 1983, *On educational testing*, San Francisco, Jossey-Bass.
- ANGOFF, W. H., 1971, Scales, normes, and equivalent scores, In: THORNDIKE, R. L., (ed.), 1972, *Educational measurement*, Washington DC: American Council of Education.
- BEJAR, I. I., 1980, Introduction to item response models and their assumptions, In: HAMBLETON, R. K. (ed.), 1983, *Applications of item response theory*, Vancouver, BC: Educational Research Institute of Brit. Columbia, 1—23.
- BIRNBAUM, A., 1968, Some latent trait models and their use in inferring an examinee's ability, In: LORD, F. M., NOVICK, M. R., 1968, *Statistical theories of mental test scores*, Reading, MA: Addison-Wesley.
- COOK, L. L., EIGNOR, D. R., 1983, Practical considerations regarding the use of item response theory to equate tests, In: HAMBLETON, R. K. (ed.), 1983, *Applications of item response theory*, Vancouver, BC: Educational Research Institute of Brit. Columbia.
- COOK, L. L., EIGNOR, D. R., 1989, Using item response theory in test score equating, *International J. of Ed. Research*, 161—174.
- CRONBACH, L. J., 1947, Test „reliability”, its meaning and determination, *Psychometrika*, 305—310.

- CRONBACH, L. J., 1951, Coefficient alpha and the internal structure of test, *Psychometrika*, 297—333.
- CRONBACH, L. J., 1984, *Essentials of psychological testing*, New York, Harper and Row.
- EMBRETSON, S. E. (ed.), 1985, *Test design, developments in psychology and psychometrics*, Orlando, San Diego.
- FISCHER, G. H., 1974, *Einführung in die Theorie psychologischer Tests*, Bern, Huber.
- FISCHER, G. H. (ed.), 1968, *Psychologische Testtheorie*, Bern, Huber.
- FISCHER, G. H., 1981, On the existence and uniqueness of maximum-likelihood estimates in the Rasch Model, *Psychometrika*, 59—77.
- FISCHER, G. H., 1987, Applying the principles of specific objectivity and of generalizability to the measurement of change, *Psychometrika*, 565—587.
- FISCHER, G. H., PARZER, P., 1991, An Extension of the Rating Scale Model with an Application to the Measurement of Change, *Psychometrika*, 1991, dec.
- FISCHER, G. H., KISSER, R., 1983, Notes on the exponential latency model and an empirical application, In: WAINER, H., MESSICK, S. (eds.), 1983, *Principles of modern psychological measurement, A festschrift for Frederic M. Lord*, Hillsdale, NJ: Lawrence Erlbaum, 139—185.
- FISCHER, R., 1959, *Statistical methods and scientific inference*, Edinburgh, Oliver and Boyd.
- FORMANN, A. K., 1984, *Die Latent-Class-Analyse, Einführung in die Theorie und Anwendung*, Weinheim, Beltz.
- FREDERICKSEN, N., GLASER, N. (eds.), 1990, *Diagnostic monitoring of skill and knowledge acquisition*, Hillsdale, NJ: Lawrence Erlbaum.
- GOLDSTEIN, H., WOOD, R., 1989, Five decades of item response modelling, *Brit. J. of Math. and Stat. Psych.*, 139—167.
- GREEN, B. F., 1983, The promise of tailored tests, In: WAINER, H., MESSICK, S. (eds.), *Principles of modern psychological measurement, A festschrift for Frederic M. Lord*, Hillsdale, NJ: Lawrence Erlbaum.
- GULLIKSEN, H., 1960, *Theory of mental tests*, New York: Wiley.
- GUTTMAN, L., 1945, A basis for analysing test-retest reliability, *Psychometrika*, 255—282.
- GUTTMAN, L., 1971, Measurement as structural theory, *Psychometrika*, 1971, Dez., (346.).
- HAMBLETON, R. K., 1989, Principles and selected applications of item response theory, In: LINN, R. L. (ed.), 1989, *Educational measurement*, New York: Macmillan, 147—200.
- HAMBLETON, R. K., 1990, Item response theory: introduction and bibliography, *Psychothema*, 97—107.
- HAMBLETON, R. K. (ed.), 1983, *Applications of item response theory*, Vancouver, BC: Educational Research Institute of Brit. Columbia.

- HAMBLETON, R. K., MURRAY, L. N., 1983, Some goodness of fit investigations for Item Response Models, In: HAMBLETON, R. K. (ed.), 1983, *Applications of item response theory*, Vancouver, BC: Educational Research Institute of Brit. Columbia.
- HAMBLETON, R. K., ROGERS, H. J., 1983, Solving criterion-referenced measurement with item response models, In: HAMBLETON, R. K. (ed.), 1983, *Applications of item response theory*, Vancouver, BC: Educational Research Institute of Brit. Columbia.
- HAMBLETON, R. K., ROGERS, H. J., 1989, Detecting potentially biased test items: comparison of area and Mantel-Haenszel methods, *Applied Measurement in Ed.*, 313—334.
- HAMBLETON, R. K., SWAMINATHAN, H., 1985, *Item Response Theory: principles and applications*, Boston, MA: Kluwer.
- HAMBLETON, R. K., ZAAL, J. N., PIETERS, J. P. M., 1991, Computerized adaptive testing: theory, applications, and standards, In: HAMBLETON, R. K., ZAAL, J. N. (eds.), 1991, *Advances in educational and psychological testing: theory and applications*, Boston, MA: Kluwer.
- HAMBLETON, R. K., ZAAL, J. N. (eds.), 1991, *Advances in educational and psychological testing: theory and applications*, Boston, MA: Kluwer.
- HORVÁTH Gy., 1985, Tesztelmélet: problémák, perspektívák, *Pszichológia*, 1, 53—78.
- HORVÁTH Gy., 1991, *Az értelem mérése*, Budapest, Tankönyvkiadó.
- HULIN, C. L., DRASGOW, F., PARSONS, C. K., 1983, *Item response theory, Application to psychological measurement*, Homewood.
- JULIAN, E. R., WRIGHT, B. D., 1988, Using computerized patient simulation to measure the clinical competence of physicians, *Applied Measurement in Ed.*, 299—318.
- KEMPF, W. F., REPP, B. H. (eds.), 1977, *Mathematical models for social psychology*, Bern: Huber and New York: Wiley.
- KLAUER, K. J., 1978, *Handbuch der Pädagogischen Diagnostik*, 1—4, Band, Düsseldorf.
- KUBINGER, K. D. (ed.), 1988, *Moderne Testtheorie, Ein Abriss samt neuesten Beiträgen*, München.
- KUDER, G. F., RICHARDSON, M. W., 1937, The theory of the estimation of test reliability, *Psychometrika*, 151—160.
- LAZARSFELD, P. F., HENRY, N. W., 1968, *Latent structure analysis*, Boston, MA: Houghton Mifflin.
- LEWIS, C., 1983, Bayesian inference for latent abilities, In: ANDERSON, S. B., HELMICK, J. S. (eds.), 1983, *On educational testing*, San Francisco: Jossey-Bass, 224—251.
- LEWIS, C., 1986, Test theory and Psychometrika: the past twenty-five years, *Psychometrika*, 11—12.
- LEWIS, C., SHEEHAN, K., 1990, Using bayesian decision theory to design a computerized mastery test, *ETS Research Report*, November, 1990.

- LIENERT, G. A., 1961, *Testaufbau und Testanalyse*, Weinheim: Julius Beltz.
- LIENERT, G. A., 1974, *Tesztyszerkesztés és tesztanalízis*, kivonat, szerk: PERCZEL T., Budapest: FPK.
- LINN, R. L. (ed.), *Educational measurement*, New York: Macmillan.
- LOEVINGER, J., 1947, A systematic approach to the construction and evaluation of tests of ability, *Psychological Monographs*, No 61.
- LOEVINGER, J., 1948, The technic of homogeneous tests compared with some aspects of „scale analysis” and factor analysis, *Psychol. Bull.*, 507—529.
- LORD, F. M., 1980, *Applications of item response theory to practical testing problems*, Hillsdale, NJ: Lawrence Erlbaum.
- LORD, F. M., 1983, Statistical bias in maximum likelihood estimators of item parameters, *Psychometrika*, 425—435.
- LORD, F. M., NOVICK, M. R., 1968, *Statistical theories of mental test scores* (with contributions of A. BIRNBAUM), Reading, MA: Addison-Wesley.
- LUNZ, M. E., WRIGHT, B. D., LINACRE, J. M., 1990, Measuring the impact of judge severity on examination scores, *Applied Measurement in Education*, 331—345.
- MASTERS, G. N., WRIGHT, B. D., 1981, Defining a fear-of-crime variable: a comparison of two Rasch Models, *Ed: Research and Perspectives*.
- MISLEVY, R. J., 1986, Bayes model estimation in Item Response Models, *Psychometrika*, 177—196.
- MISLEVY, R. J., 1989, Foundations of a new test theory, ETS Research Report, October, 1989.
- MISLEVY, R. J., VERHELST, N., 1991, Modeling item responses when different subjects employ different solution strategies, *Psychometrika*, 1990, 195—215.
- MOKKEN, R. J., 1971, *A theory and procedure of scale analysis: with applications in political research*, The Hague: Mouton and Co.
- MOKKEN, R. J., LEWIS, C., 1982, A nonparametric approach to the analysis of dichotomous item responses, *Applied Psych. Measurement*, 417—430.
- MCDONALD, R. P., 1983, Future directions for item response theory, In: HAMBLETON, R. K. (ed.), 1983, *Applications of item response theory*, Vancouver, BC: Educational Research Institute of Brit. Columbia, 205—217.
- NOVICK, M. R., 1980, Statistics as psychometrics, *Psychometrika*, 411—424.
- PERLINE, R., WRIGHT, B. D., WAINER, H., 1979, The Rasch Model as additive conjoint measurement, *Applied Psych. Measurement*, 237—255.
- RASCH, G., 1960, *Probabilistic models for some intelligence and attainment tests*, Copenhagen: Danish Institute for Educational Research.
- RICHARDSON, M. W., 1936, The relation between the difficulty and the differential validity of a test, *Psychometrika*, No 2, 33—75.
- ROSKAM, E. (ed.), 1985, *Measurement and personality assessment*, North-Holland.

- SHEEHAN, D., MISLEVY, R. J., 1990, Cognitive and Psychometric Models to Measure Document Literacy, *J. of Ed. Measurement*, No 3.
- SPADA, H., KEMPF, W. F. (eds.), 1977, *Structural models of thinking and learning*, Bern: Huber.
- STONE, M. H., WRIGHT, B. D., 1983, Measuring attending behavior and short-term memory with Knox's cube test, *Ed. and Psych. Measurement*, 803—815.
- SUEN, H. K., 1990, *Principles of test theories*, Hillsdale, NJ: Lawrence Erlbaum.
- SWAMINATHAN, H., 1983, Parameter estimation in item response models, In: HAMBLETON, R. K. (ed.), 1983, *Applications of item response theory*, Vancouver, BC: Educational Research Institute of Brit. Columbia.
- TATSUOKA, K. K., 1990, Toward an integration of item-response theory and cognitive error diagnosis, In: FREDERICKSEN, N., GLASER, N. (eds.), 1990, *Diagnostic monitoring of skill and knowledge acquisition*, Hillsdale, NJ: Lawrence Erlbaum.
- TATSUOKA, K. K., BIRENBAUM, M., ARNOLD, J., 1989, On the stability of students rules of operation for solving arithmetic problems, *J. of Ed. Measurement*, 351—361.
- TATSUOKA, K. K., LINN, R. L., TATSUOKA, M. M., YAMAMOTO, K., 1988, Differential item functioning resulting from the use of different solution strategies, *J. of Ed. Measurement*, 301—319.
- THORNDIKE, E. L., 1972, *The measurement of intelligence*, New York, Columbia University.
- THORNDIKE, R. L. (ed.), 1972 *Educational measurement*, Washington DC: American Council of Education.
- THURSTONE, L. L., CHAVE, E. J., 1929, *The measurement of attitude; a psychophysical method and some experiments with a scale for measuring attitude toward the church*, Chicago, The Univ. of Chicago Press.
- TUCKER, L. R., 1946, Maximum validity of a test with equivalent items, *Psychometrika*, 1—3.
- WAINER, H., MESSICK, S. (eds.), 1983, *Principles of modern psychological measurement, A festschrift for Frederic M. Lord*, Hillsdale, NJ: Lawrence Erlbaum.
- WAINER, H., WRIGHT, B. D., 1980, Robust estimation of ability in the Rasch Model, *Psychometrika*, 373—391.
- WEISS, D. J., KINGSBURY, G. G., 1984, Application of computerized adaptive testing to educational problems, *J. of Ed. Measurement*, 361—375.
- WILD, B., 1988, Neue Erkenntnisse zur Effizienz des „tailored“ — adaptiven Tests, In: KUBINGER, K. D. (ed.), 1988, *Moderne Testtheorie, Ein Abriss samt neuesten Beiträgen*, München.
- WILD, B., 1988, Neue Simulationsstudien zur Effizienz verschiedener Paramterschatz und Itemauswahl Strategien, In: KUBINGER, K. D. (ed.),

- 1988, *Moderne Testtheorie, Ein Abriss samt neuesten Beiträgen*, München.
- WILLMOTT, A. S., FOWLES, D. E., 1974, *The objective interpretation of test performance*, The Rasch Model applied, Windsor.
- WOOD, R. L., WINGERSKY, M. S., LORD, F. M., 1976, Logist, A computer program for estimating examinee ability and item characteristic curve parameters, Princeton, NJ: ETS.
- WRIGHT, B. D., PANCHAPAKESAN, N., 1969, A procedure for sample-free item-analysis, *Ed. and Psych. Measurement*, 23—48.
- WRIGHT, B., MASTERS, J., 1982, *Rating scale analysis*, Chicago, MESA Press.
- WRIGHT, B., 1988, *Some fundamentals of measurement*, Chicago, kézirát.
- WRIGHT, B., STONE, M., 1979, *Identification of item bias*, Chicago, kézirát.
- WRIGHT, B., STONE, M., 1979, *Best test design*, Chicago, MESA Press.
- WRIGHT, B. D., 1977, Solving measurement problems with the Rasch Model, *J. of. Ed. Measurement*, 97—116.
- WRIGHT, B. D., 1985, Additivity in psychological measurement, In: ROSKAM, E. (ed.), 1985, *Measurement and personality assessment*, North-Holland.
- WRIGHT, B. D., DOUGLAS, G. A., 1977, Best procedures for sample-free item analysis, *Applied Psych. Measurement*, 281—295.
- WRIGHT, B. D., DOUGLAS, G. A., 1977, Conditional versus unconditional procedures for sample-free item analysis, *Ed. and psych. Measurement*, 573—586.
- WRIGHT, B. D., BELL, S. R., 1984, Item banks: what, why, how, *J. of ed. Measurement*, 331—345.
- YAMAMOTO, K., 1989, Hybrid model of IRT and latent class models, *ETS Research Report*, September, 1989.