Pszichológia, 1985, (5), 1, 53–78

HORVÁTH GYÖRGY
Budapesti Műszaki Egyetem

TESZTELMÉLET: PROBLÉMÁK ÉS PERSPEKTÍVÁK

„Klasszikus" tesztelmélet – ezzel a szóhasználattal sűrűn találkozhatunk a (viszonylag) újabb keletű (az utóbbi egy-két évtizedben íródott) tesztelméleti tanulmányokban. A „klasszikus" jelző ellenpárja ilyenkor a „*modern*" vagy a „*probabilisztikus*". A modern tesztelmélet megjelölés egyre inkább a különféle valószínűségi szemléletű – sztochasztikus, probabilisztikus – modellekre vonatkozik, szemben a korábbi tesztelmélet, úgymond, determinisztikus felfogásával.

A klasszikus tesztelmélet csúcsának, betetőzésének általában GULLIKSEN 1950-ben és LORD és NOVICK 1968-ban megjelent kötetét tartják: ezekben a munkákban a klasszikus tesztelmélet axiomatikus formát öltött. Ezzel egyszersmind jól láthatóvá, tudatossá váltak korlátai is: az *alapfeltevések*, amelyekre épít, és amelyek teljesülése megkérdőjelezhető.

A korlátok meghaladására irányuló kísérletek azonban már korábbi keletűek, s a szociálpszichológia-szociológia határterületeiről indultak el: az attitűdvizsgálatokhoz kapcsolódva alakult ki még a II. világháború idején a „latent structure" (látens struktúra) irányzat, amely főként Paul Lazarsfeld nevéhez fűződik. Jelentős lépés volt. L. Guttmann ún. „perfect scale"-ja (perfekt skálája) (sajátos attitűdskála-szerkesztési eljárása), amely még determinisztikus jellegű ugyan, de újszerű elméleti meggondolásokra és termékenynek bizonyult fogalmakra épült. Az ötvenes évek vége felé és a hatvanas években indult meg a szűkebb értelemben vett tesztvizsgálatokban is a különböző probabilisztikus modellek kidolgozása. A princetoni teszteoretikusok egy csoportja (L.R. Tucker, F. Lord, A. Birnbaum) úttörő szerepet játszott az új szemléletmód megalapozásában, és a Lord és Novick-féle kézikönyv már közölhette A. Birnbaum áttekintését a probabilisztikus tesztelmélet első lépéseiről. Tőlük függetlenül született a dán RASCH (1960) munkája, mely a többi között a dán hadsereg részére kifejlesztett intelligenciateszt metodológiai megalapozását ismertette. A Rasch-modell a modern tesztelméleten belül immár a klasszikus rangjára tett szert, és mindmáig a leggyakrabban emlegetett mintája és kiindulópontja az újabb tesztelméleti fejlődésnek.

A modern tesztelmélet első jelentős összefoglalása azonban a bécsi egyetem professzorának, FISCHERnek (1974) a nevéhez fűződik. A bécsi egyetem azóta is a modern tesztelméleti kutatások egyik legjelentősebb műhelye és – lévén a német nyelvterület legnagyobb pszichológusképző központja (jelenleg a pszichológushallga-

tók létszáma a bécsi egyetemen mintegy háromezer fő!) – egyúttal az új szemléletmód hatékony terjesztője is.*

A klasszikus tesztelmélet alapfeltevései és korlátai

A klasszikus tesztelmélet bírálói mindenekelőtt arra hívják fel a figyelmet, hogy ez az „elmélet" paradox módon nem igazából elmélet: éppen elméletileg nem kellően megalapozott. Ami abban mutatkozik meg, hogy a klasszikus teszttan főként a tesztszerkesztés statisztikai-technikai eljárásainak pragmatikus gyűjteménye, fel sem veti azonban a pszichológiai értelmezés számára döntő jelentőségű, alapvető méréselméleti problémákat. Holott egy valódi elméletnek meg kellene előznie – meg kellene alapoznia – a tesztpontértékek megállapítását, a tulajdonképpeni „mérést". Ezzel szemben a klasszikus teszttan kiindulásában a tesztek által szolgáltatott pontértékeket mintegy adottnak veszi, valamely a priori létező mennyiség adekvát mérőszámainak tekinti, és a továbbiakban csak arra összpontosítja figyelmét, hogy célszerű módon miként kezelhetők ezek a mérőszámok.

A klasszikus teszttan szemléleti kiindulópontja – tehát az, hogy a tesztpontértékeket egy valóságosan létező mennyiség közvetlen mérőszámaiként fogadja el – e tesztelmélet alapegyenletében kristályosodik ki:

Mért tesztpontérték *(X)*= Valódi érték *(T)* + Mérési hiba *(E)*

A tesztpontértékek ezek szerint a *mérési hibától eltekintve* közvetlenül megfeleltethetők az egyértelműen meghatározott „valódi értékeknek" (true score, wahrer Wert, isztyinnij ball), amelyek létezése az elmélet legfontosabb alapfeltevése. Az ún. tesztelméleti axiómák is a mérési hibák és a valódi értékek viszonyára vonatkozó tételeket (előfeltevéseket) mondanak ki. Ezek szerint: *a)* a mérési hibák várható értéke zéró; *b)* a valódi érték és a mérési hiba közti korreláció zéró; *c)* két különböző teszt (illetve tesztitem) esetében az egyik valódi értéke és a másik mérési eredménye közti korreláció zéró; *d)* két különböző tesztben, (illetve tesztitemben) a mérési hibák korrelációja zéró. Így válik a továbbiakban a mérési eljárás – az előfeltevéseknek megfelelően, elvben – a fizikai mérésekhez hasonlatossá. Az intelligencia ekkor úgy mérhető, akár a testsúly: kisebb-nagyobb mérési hibával, a mért eredmény azonban egy „valódi érték" körül ingadozik. Az elmélet – az alkalmazott statisztikai eljárások – feladata főként a „valódi érték" és a mérési hibák viszonyának leírása, a mérési hiba becslése, csökkentése, a mérés megbízhatóságának fokozása. A klasszikus tesztelmélet központi kérdésköre ezért a mérési pontosságnak (reliabilitásnak) és az előrejelzés pontosságának (a validitásnak) a problémája.

A fenti előfeltevésekből a varianciákra vonatkozóan is levezethető egy összefüggés, mely a reliabilitás fogalmához is elvezet:

*1982 tavaszán a Collegium Hungaricum ösztöndíjasaként személyesen is megismerkedhettem a bécsi egyetem pszichológiai intézetének munkájával, az ott folyó elméleti és gyakorlati kutatásokkal. Konzultálhattam az intézet igazgatójával, G.H. Fischerrel és munkatársaival, A.K. Formannal, B. Rollettal és másokkal, akik a legnagyobb készséggel álltak rendelkezésemre a szakmai tájékozódásban. Ehelyütt is köszönöm segítségüket, amely e cikk megírásához is hozzájárult.

Mért tesztpontértékek varianciája (σ_X^2)= Valódi értékek varianciája (σ_T^2) + Hibavariancia (σ_E^2).

Ennek alapján valamely teszt reliabilitásának elméleti definíciója a következő:

$$\text{Reliabilitás} = \frac{\text{Valódi értékek varianciája } (\sigma_T^2)}{\text{Mért értékek varianciája } (\sigma_X^2)} = 1 - \frac{\text{Hibavariancia } (\sigma_E^2)}{\text{Mért értékek varianciája } (\sigma_X^2)}$$

A reliabilitás tehát a varianciák aránya révén fejezi ki a mérés pontosságát; megadja a „valódi" varianciának a teljes varianciában való százalékos részesedését. Ha a mérés hibáktól mentes (azaz a hibavariancia zéró), akkor az első tört (az ún. reliabilitás-koefficiens) értéke 1, ha viszont a mérés teljesen hibás, akkor a tört értéke zéró.

Az előfeltevések felhasználásával megmutatható az is, hogy a varianciák arányaként definiált reliabilitás felírható a mért értékek és a valódi értékek (a mért dimenzió) közti korreláció négyzeteként, amely a reliabilitás fenti definíciójával egyenértékű definíciónak tekinthető. (A reliabilitás-koefficienstől való megkülönböztetésül a reliabilitás-koefficiens négyzetgyökét, vagyis az említett korrelációs együtthatót reliabilitásindexnek nevezik.)

Mindkét definíció elméleti jellegű: gyakorlati számításokhoz nem használhatók fel, mivel a „valódi értékek" (illetve varianciájuk) nem ismeretesek. (A „valódi értékeket" — definíció szerint — úgy kaphatnánk meg, ha végtelen számú párhuzamos mérés átlagát vennénk.) Tekintsünk most el a gyakorlati számítások nehézségeitől, a további, pszichológiai szempontból kevéssé realisztikus feltevéseknek (a tesztek paralelitásának) problematikus voltától. Az elméleti definíciókból is nagyon fontos tanulság adódik.

Ha ugyanis a mért értékek varianciája (σ_X^2) valamely okból megváltozik, akkor szükségképpen megváltozik a reliabilitás értéke is; a reliabilitásértékben a csoport homogén vagy heterogén volta is tükröződik.

Vagyis a reliabilitás nem egyedül a szóban forgó tesztnek, hanem a tesztnek és a figyelembe vett populációnak *együttes* jellemzője. Jól szemlélteti ezt az egyszerű gondolatkísérlet: tekintsünk egy végletesen homogén populációt, amelyben mindenki ugyanazzal a „valódi értékkel" rendelkezik. Ekkor a „valódi variancia" és vele együtt a reliabilitás is zéró lesz. Úgy tűnik, hogy a teszt minden mérési pontosságát elvesztette. Tulajdonképpen azonban az történt, hogy az adott populációba tartozó egyének a teszt révén többé nem különböztethetők meg egymástól. Ami mutatja, hogy a reliabilitás valójában annak a mértéke, mennyire különböztethetők meg a populációba tartozó egyének az X változó alapján. Tehát: „A reliabilitás a teszt pontosságát egy adott populáció vonatkozásában méri." (FISCHER, 1974, 38. o.)

A tesztjellemzőknek ezt az alapvető sajátosságát —, hogy a mindenkori referencia populációtól függnek — a populációtól való függőségnek nevezik (a probabilisztikus tesztek „populációfüggetlensége" pedig az ebben az értelemben vett *populációfüggőségtől* való mentességet, a mintapopuláció összetételével szembeni invarianciát jelenti). A validitás (amely a teszteredményeknek egy kritériumváltozóval való korrelációjaként definiálható, s amelynek így a reliabilitás csak sajátos esete) ugyancsak populációfüggő. (Mint ahogy populációfüggő minden hasonló jellegű, varianciák arányában kifejeződő mutató is, így populációfüggőek a genetikusok által használt heritabilitásmutatók is!) A populációfüggőség paradox következményei különösen

szembetűnőek a klasszikus tesztelmélet fejlődésének csúcsait képviselő faktoranalízisben. Nagy János tanuló intelligenciájának (képességeinek, személyiségének stb.) faktorstruktúrája más és más lesz aszerint, hogy melyik populáció tagjaként szerepeltetjük: az első gimnazisták populációjába sorolódott, vagy csak a fiútanulók populációjának tagjaként vizsgálták, netán a sportolók felmérésébe került bele stb. A faktorstruktúráknak ez a (figyelembe vett populációtól függő) különbözősége szöges ellentétben van a faktoranalízis kiinduló alapfeltevéseivel.

Eddig nem érintettük a *skálaproblémákat.* Hallgatólagosan tudomásul vettük, hogy a tesztelmélet mind a „valódi értékek", mind pedig az ezeknek (mérési hibáktól eltekintve) közvetlenül megfelelő tesztpontértékek vonatkozásában legalább intervallumskála létezését feltételezi. Hiszen enélkül nem szabad, nem lehet varianciát, de még csak átlagot sem számolni.

Ámde ez a hallgatólagos feltételezés, miként azt F. LORD (1953) megmutatta, szintén paradox eredményekhez, ellentmondáshoz vezet. Ha ugyanazon populációnak két különböző nehézségű (egy túl könnyű és egy túl nehéz) tesztet adunk, a tesztpontértékek eloszlása eltérő lesz; a könnyebb tesztet többen oldják meg, az eloszlás a magasabb pontértékek irányában lesz aszimmetrikus, a nehezebb teszt esetében az eltérés fordított irányú. A két eltérő eloszlás egyidejűleg *ugyanazon* megengedett transzformációval nem hozható közös „normális" alakra. Semmilyen alapunk sincs rá, hogy a két eltérő eloszlást eredményező teszt közül valamelyiknek kitüntetett szerepet tulajdonítsunk, tehát el kell fogadnunk, hogy a két teszt eltérő valódiérték-skálákhoz vezet. Ha feltételezzük (s e feltételezés nélkül a mérés eleve értelmetlen volna), hogy maga a mérendő (de a közvetlenül nem mérhető, látens) tulajdonság (ability score) egyenlő nagyságú egységekkel (intervallumskálán) jellemezhető, akkor szükségképpen következik, hogy sem a nyers pontértékek, sem a „valódi értékek" skáláján nem egyenlő közűek a beosztások (LORD, 1973, 68. o.). Eysenck a teszteredmények itemfüggőségén alapuló kritikával — tehát a Lord által is felhasznált érveléssel — szemben ellenérvként azt hozza fel (EYSENCK, 1980, 72. o.), hogy a tesztszerkesztés szabályai előírják a különféle nehézségi szintű itemek lehetőleg egyenletes elosztását, és ez a pontértékek közelítőleg normális eloszlásához kell, hogy vezessen. Ámde a nehézségi szint ugyancsak populációfüggő, csakis valamely adott populációra vonatkoztatva határozható meg. Az itemfüggőség csupán a populációfüggőség másik oldala.

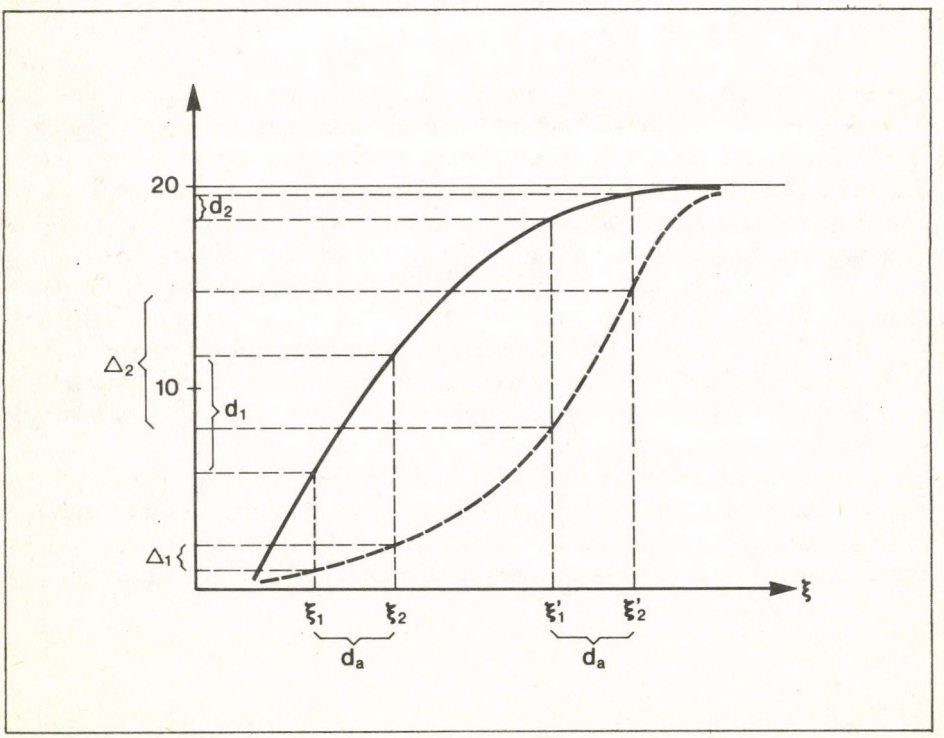
A Lord által elemzett ellentmondásoktól függetlenül is tény, hogy a klasszikus tesztelmélet az intervallumskála-tulajdonságok meglétét egyszerűen *feltételezi* minden további külön megfontolás nélkül. Ugyanez vonatkozik a normális eloszlás *feltevésére* is (lásd bővebben GUTJAHR, 1974, 164. o.), amellyel gyakorta az előző, az intervallum-skálára vonatkozó feltevést próbálják indokolni. Ezzel kapcsolatban G.H. Fischer alapos méréselméleti fejtegetéseit azzal az összefoglaló megállapítással zárja, hogy „teljesen hamis a régi keletű elképzelés, amely szerint standardizálás révén (tehát egy meghatározott populáció teszteredményeinek normális eloszlássá transzformálásával) bármit is eldönthetnénk a skálatulajdonságokra vonatkozóan. Eloszlásnak és skálának okozatilag semmi közük egymáshoz, akkor sem, ha az egyik mesterséges változtatása a másik változtatását vonja is maga után" (FISCHER, 1974, 125. o.).

Nem térhetünk itt ki a klasszikus tesztelmélet kritikájának minden vonatkozására, például a homogenitás- és ekvivalencia-problémákra, vagy a validitásnak a reliabili-

tásnövekedéssel párhuzamos „sorvadására", az ún. „attenuation"-paradoxonra). Ezért még csak a *változásmérések* közismerten ingoványos terepére utalunk, minthogy a kísérletező pszichológus, a pedagógiai pszichológia vagy a fejlődéslélektan kutatója különösen gyakran találja magát szembe a változásmérés feladatával. Ilse Ropnak, a bécsi egyetem kutatójának tanulságos vizsgálatai, melyeket ugyancsak Fischer ismertet, az itt szükséges óvatosságra intenek.

Legyen egy óvodai felzárkóztató program hatásvizsgálata a feladat. Minthogy a program célja az esélyegyenlőség előmozdítása, nemcsak az a kérdés, hogy a felzárkóztató program a kontrollcsoport munkájához képest átlagosan javulást eredményez-e, hanem az is, hogy a csoporton belül kik profitálnak belőle erősebben, a hátrányos helyzetű, gyengébb gyerekek, vagy a már amúgy is kedvezőbb helyzetben levő fejlettebbek.

Ez a kérdés a klasszikus tesztelmélet módszereivel nem válaszolható meg egyértelműen. Az (idealizált) *1. ábra* szemléleti, hogy a tesztitemek nehézségi fokától füg-



1. ábra
A ξ intelligenciafaktor fejlődésének és az *X* teszteljesítménynek összefüggése két eltérő környezetből származó gyereknél (FISCHER, 1974, 130. o.)

gően egymással ellentétes következtetésekre is juthatunk. A ξ-tengely egy látens teljesítménydimenziót, például az intelligenciafaktort, az ordináta pedig az éppen ezt a

faktort mérő tesztben felmutatott teljesítményeket tünteti fel. A folytonos görbe az intelligenciafaktor „fejlettsége" és az elért teszteredmények közti (általában nem lineáris) összefüggést mutatja. Legyen ξ_1 egy hátrányos helyzetű, ξ_2 egy kedvezőbb helyzetű gyerek „fejlettsége" a ξ faktort tekintve. Ekkor a kedvezőbb helyzetű gyerek előnye a hátrányos helyzetűvel szemben $d_a = \xi_2 - \xi_1$. Tegyük fel, hogy az öthónapos felzárkóztató program mindkettőjükre ugyanakkora hatással volt (és természetes fejlődésük is hasonló ütemben zajlott). Ekkor a d_a különbség változatlan maradt, csak a ξ-tengely mentén eltolódott.

Mit mutat azonban a teszt? Attól függ. Például — miként az ábráról leolvasható — azt is mutathatja, hogy a két gyerek közti különbség csökkent ($d_2 < d_1$), a felzárkóztatás sikeres volt. Ha a jobbik gyerek már az első felméréskor viszonylag magas teljesítményt nyújtott, most nehezebben javíthat („tetőeffektus"), az alacsonyabb teljesítményből kiindulva viszont sokkal könnyebb a javítás. Más, pontosan ellenkező eredményhez jutunk, ha a tesztet azáltal nehezítjük, hogy a könnyebb itemek egy részét elhagyjuk. Ekkor előfordulhat (lásd a szaggatott görbét), hogy éppen az eleve jobb gyerek erőteljesebb fejlődését ($\Delta_2 > \Delta_1$) kapjuk eredményül — noha a ξ látens dimenziót tekintve az eredeti különbség sem nem nőtt, sem nem csökkent.

Nyilvánvalóan jogos követelmény bármely teszttel szemben, hogy a tesztitemek kiválasztása ne vezethessen az eredmények szisztematikus torzításához, vagyis az eredmények lényegileg ne függjenek attól a körülménytől, hogy homogén (ugyanazt a faktort, illetve látens dimenziót mérő) itemek összességéből konkrétan mely itemek szerepelnek a tesztben. A klasszikus tesztelmélet keretei közt e követelmény teljesülése nem biztosítható. A probabilisztikus tesztelméletek egyik törekvése, hogy a látens dimenzió megfelelő definiálása után a tesztpontértékek különbségei helyett a látens dimenzió mentén elfoglalt helyek ($\hat{\xi}_1 - \hat{\xi}_2$) különbségének mérésére térjen át, és ezáltal a hasonló torzításokat elkerülje.

A klasszikus tesztelmélet kritikájából természetesen nem következik annak elvetése. Ellenkezőleg: érvényességi körének behatárolása, korlátainak tudatosítása biztonságosabbá — tudományosan megalapozottabbá — teszi a tesztmódszerek gyakorlati alkalmazását mindenütt, ahol ez torzításokhoz nem vezet. A klasszikus tesztelmélet bírálata nem idegen eredetű, kívülről jövő bírálat, hanem a belső fejlődésnek, a klaszszikus tesztelméleten belül néhány évtized alatt felhalmozódott óriási tapasztalati és ismeretanyagnak a terméke. Hiszen a klasszikus tesztelméletnek is természetes törekvése a torzítások elkerülése és érvényességi körének minél pontosabb meghatározása.

Vétségek a klasszikus tesztelmélet szabályai ellen

A klasszikus tesztelméleten belül maradni annyit tesz, mint elfogadni azt a matematikai modellt, amelyre az épül, éspedig következményeivel együtt. A mindennapi gyakorlatban gyakran előfordul azonban, hogy a tesztmetodika alkalmazói nem számolnak azokkal a következményekkel, amelyek éppen a klasszikus tesztelmélet elfogadásából származnak. Néhány példát lássunk az így adódó hibalehetőségekre.

Nem mindig számolnak az alkalmazók a *tesztnormáknak* a sztenderdizálás módjából adódó *relativitásával.* Az életkori IQ helyébe lépett deviációs IQ-számítás sem

eléggé egyértelmű ugyanis: a nyert értékek értelmezése az *SD* (standard deviáció, szórás) értékétől függ, ez pedig különböző tesztek és eltérő (például életkorban eltérő) populációk esetén nem ugyanaz. Bár terjedőben van az intelligenciatesztek olyan kalibrálása, ahol az SD $\approx$ 15, azaz az egy szórásnyi intervallumnak nagyjából 15 pontértéket feleltetnek meg, de ez nem kötelező (összetett teszteknél egységesen nem is biztosítható), és egyelőre a különböző skálabeosztások tarka változatosságával találkozhatunk. A Binet-Stanford-teszt 16 pontnyi szórásértéke pedig (amelyhez a 15 pontnyi SD-t alapul vevő Wechsler-skála is igazodik) eredetileg az intelligenciahányados különböző (12 és 20 pont közti) életkori szórásértékeinek átlagaként adódott, de az átlagolás természetesen nem szüntethette meg a részpopulációk homogenitásában és a teljesítményekben az életkori eltéréseket (ami a szórás eltéréseiben fejeződik ki).

A magyar Wechsler-teszt is jól szemlélteti (lásd az *1. táblázatot*) azt az általános

1. táblázat

Az értékpontok középértékei és szórásai 10 és 59 évek között a Wechsler-teszt magyar változatában

Kor	Számolt középérték	Számolt szórás	Simított középérték	Simított szórás
10	82	16	82	16
11	85	13,1	88	15,5
12	93	15,6	93	14,8
13	102	14,3	98	15,16
14	105	15,0	103	16,22
15,5	109	17,8	107	16,9
17,5	107	18,4	108	18,48
19,5	114	19,0	109	19,66
22	109	22,2	108	20,22
25	108	20,9	108	20,52
28	106	20,6	105	21,24
31	103	19,9	103	20,92
34	100	22,6	102	20,78
37	102	20,6	100	20,84
40	99	20,2	99	21,06
43	100	20,9	98	21,12
46	98	20,9	97	21,64
49	95	23,0	95	22,82
52	93	23,2	92	23,7
55	92	26,1	88	24,6
58	84	25,6	85	25,6

(KUN és SZEGEDI, 1971, 123. o.)

tapasztalatot, hogy „az IQ-táblázat szórási értékei az egységül szolgáló életkoroktól függenek. A 10 évesek vagy 19–20 évesek korcsoportjára számolt szórásérték például egymástól különbözik, korcsoportról korcsoportra változik, és ezáltal a számított átlagtól eltér" (KUN és SZEGEDI, 1971, 131. o.) – amit a megfelelő transzformációval csak elfedni lehet. Az ún. „culture-fair" (a kulturális közeg iránt „elfogulatlan"), a „fluid intelligence"-t (a „képlékeny" intelligenciát) mérni szándékozó tesztekben pedig 20–24 körüli érték szokott lenni a szórás (a populációtól függően!).

Teljesen helyénvaló ezért Eysenck figyelmeztetése: „Ha IQ-kat hasonlítunk össze, mindig figyelembe kell vennünk az SD-k eltéréseit, különben az összehasonlítás értelmetlen" (EYSENCK, 1980, 70. o.). Az olyasfajta kijelentések tehát, hogy „az emberek mintegy kétharmadának értelmi szintje 85 és 115 IQ között van", nem korrektek, ha nem adják meg, mely teszt alapján számolják az IQ-értékeket (hogy a hasonló megállapítások egyéb problematikus mozzanatait itt ne említsük). Anastasi fontosnak tartja, hogy táblázatban is szemléltesse a forgalomban levő tesztek eltérő SD-értékeiből származó besorolási eltéréseket (*2. táblázat*; ANASTASI, 1976, 87. o.).

2. táblázat

Az IQ-értékek jelentése az SD függvényében

IQ-intervallum	Százalékos gyakoriság			
	SD= 12	SD= 14	SD= 16	SD= 18
130 és afölött	0,7	1,6	3,1	5,1
129–129	4,3	6,3	7,5	8,5
110–119	15,2	16,0	15,8	15,4
100–109	29,8 } 59,6	26,1 } 52,2	23,6 } 47,2	21,0 } 42,0
90– 99	29,8	26,1	23,6	21,0
80– 89	15,2	16,0	15,8	15,4
70– 79	4,3	6,3	7,5	8,5
70 alatt	0,7	1,6	3,1	5,1
Összesen	100,0	100,0	100,0	100,0

(ANASTASI, 1976, 87. o.)

Hogy a besorolási problémáknak milyen gyakorlati vonatkozásai és gyakorlati következményei lehetnek, arra hazai példát találunk LADÁNYI és CSANÁDY munkájában (1983, 115–120. o.).

Nem mindig számolnak (sőt nagyon ritkán számolnak) a tesztek alkalmazói a *tesztpontérték-különbségek értelmezésekor* a mérési hibákból szükségképpen adódó bizonytalanságokkal. Pedig a klasszikus tesztelmélet alapegyenlete szerint – mint arról

már szó volt – a mért tesztpontérték a „valódi érték" és a mérési hiba összeadódásából származik. Statisztikai meggondolásokból meghatározható az a megbízhatósági tartomány (konfidenciaintervallum), amelyen belül kellően nagy valószínűséggel megtalálható a mért pontértékből becsülhető „valódi érték". E tartomány nagysága – a mérés pontossága – nyilvánvalóan függ a mérési hiba várható nagyságától, amit a tesztpontértékek szórása *(SD)* és a reliabilitás ismeretében becsülhetünk meg a következő formula szerint: $SH = SD \cdot \sqrt{1 - R}$ (ahol *R*-rel a reliabilitást, *SH*-val a standard mérési hibát – a σ_E-t – jelöltük). Egy tipikus számpélda: ha az *SD* = 15 és a reliabilitás-koefficiens értéke 0,89, akkor az IQ pontérték standard hibájaként 5 pont adódik. Ez azt jelenti, hogy ha Nagy János tanulónál az intelligenciamérés eredményül 110 IQ-t adott, akkor nagy valószínűséggel (68%-os biztonsággal) annyit feltételezhetünk, hogy Nagy János intelligenciájának „valódi értéke" 105 és 115 között van valahol (nem pont középen!). Nagyobb (99%-os) biztonsággal csak annyit állapíthatunk meg, hogy a „valódi érték" valahol 97 és 123 IQ között helyezkedik el.

Mondhatom-e nyugodt lelkiismerettel, hogy Nagy János intelligensebb, mint Kiss József, aki ugyanabban a tesztben csak 105 IQ teljesítményt ért el? A klasszikus tesztelmélet értelmében *nem*. Hiszen a fentiekhez hasonlóan Kiss Józsefről csak azt tudom, hogy intelligenciájának „valódi értéke" valahol 100 és 110 IQ, illetve (nagyobb, 99%-os biztonsággal) 92 és 118 IQ között van. Előfordulhat tehát az is, hogy Kiss József IQ-jának „valódi értéke" nem kevesebb, mint 20 ponttal jobb Nagy Jánosénál (Guilford éppen azért vezette be a durvább, összesen tíz intervallummal dolgozó stanine-skálát, mert a túl sűrűn kalibrált IQ-skála látszólagos pontossága félrevezető!) Az *SD* és a reliabilitás ismeretében kiszámíthatjuk az ún. kritikus differenciát (LIENERT, 1961, 446–447. o.): mennyivel kell legalább nagyobbnak lennie az egyik mért IQ-nak a másiknál, hogy kellően nagy valószínűséggel a „valódi értékek" eltérésére következtethessünk (lásd a *3. táblázatot*). (A „valódi értékek" eltérésének mértékére természetesen csak a mérési hiba által megengedett pontossággal nyerhetünk információt, különbségek mérésekor pedig a mérési hibák összeadódhatnak. Bonyolítja továbbá a helyzetet, hogy a reliabilitás populációfüggő, ezért nem volna jogosult egyszerűen kölcsön venni reliabilitásmutatót, miként azt a magyar Wechsler-változat esetében – ahol a standardizáláskor reliabilitást nem számoltak – kénytelenek vagyunk megtenni.)

Hasonló problémák adódnak, ha a teszteredmények stabilitására, például az IQ-értékek időbeli változásaira vagyunk kíváncsiak. S ezzel eljutottunk a pszichológiai-pedagógiai kísérletezésben oly gyakori változásmérésekhez, melyek nehézségeivel ugyancsak nem mindig (hanem csak ritkán) számolunk, holott azt, aki „a pszichológiai adatok feldolgozásával kapcsolatosan már első naiv ártatlanságát elvesztette, fásultság fogja el, valahányszor különbségek méréséről hall", annyi a buktató (és a bukás) a változásmérésekben (STELZL, 1982, 184. o.): a differenciák skálatranszformációkkal manipulálhatók, többnyire nem megbízhatóak, a differenciákkal számított korrelációk pedig igen gyakran művileg előállított tényekhez, áltényekhez vezetnek. Egy példát – az óvodai felzárkóztató program hatásvizsgálatával kapcsolatban – már idéztünk a hibalehetőségek szemléltetésére. De Stelzl oly bőséggel mutat be további konkrét szakirodalmi példákat (igen kiváló szerzők munkásságából is) a tesztelmélet statisztikai

3. táblázat

A kritikus differencia a reliabilitás és az *SD* függvényében (95%-os megbízhatósági szinten)

Reliabilitás	Kritikus differencia								
	SD=10	*SD*=12	*SD*=14	*SD*=15	*SD*=16	*SD*=18	*SD*=20	*SD*=22	*SD*=24
0,85	7,59	9,11	10,63	11,39	12,15	13,66	15,18	16,70	18,22
0,86	7,33	8,80	10,27	11,00	11,73	13,20	14,67	16,13	17,60
0,87	7,07	8,48	9,89	10,60	11,31	12,72	14,13	15,55	16,96
0,88	6,79	8,15	9,51	10,18	10,86	12,22	13,58	14,94	16,30
0,89	6,50	7,80	9,10	9,75	10,40	11,70	13,00	14,30	15,60
0,90	6,20	7,44	8,68	9,30	9,92	11,16	12,40	13,64	14,88
0,91	5,88	7,06	8,23	8,82	9,41	10,58	11,76	12,94	14,11
0,92	5,54	6,65	7,76	8,32	8,87	9,98	11,09	12,20	13,30
0,93	5,19	6,22	7,26	7,78	8,30	9,33	10,37	11,41	12,45
0,94	4,80	5,76	6,72	7,20	7,68	8,64	9,60	10,56	11,52
0,95	4,38	5,26	6,14	6,57	7,01	7,89	8,77	9,64	10,52
0,96	3,92	4,70	5,49	5,88	6,27	7,06	7,84	8,62	9,41
0,97	3,39	4,07	4,75	5,09	5,43	6,11	6,79	7,47	8,15
0,98	2,77	3,33	3,88	4,16	4,43	4,99	5,54	6,10	6,65
0,99	1,96	2,35	2,74	2,94	3,14	3,53	3,92	4,31	4,70

szabályaival szembeni vétségekre, az így előállított hamis eredményekre, hogy fásult sóhaja valóban indokoltnak tűnik.

Ezért csupán még egy hibalehetőségre, gyakori vétségre utalunk: *különböző populációk* IQ-jának *összehasonlítására*. Ezt nem annyira a pszichológusok, mint inkább egyéb szakemberek szokták elkövetni, minthogy számukra az intelligenciatesztek normáinak, paramétereinek populációfüggősége és kulturális hatásoktól való függése, továbbá a tesztalkalmazásnak a tesztelmélet által is körvonalazott korlátai kevésbé nyilvánvalóak. (Érdekes, hogy noha a genetikusok hivatalból tisztában vannak a heritabilitás-mutatók *populációfüggőségével*, mihelyst az intelligenciáról van szó, erről is szívesen megfeledkeznek.) A tesztpontértékeknek, IQ-értékeknek a mérőeszközöktől való függetlenítése, elvont mennyiségként való kezelése, túlértékelése (Székely Lajos kifejezésével: a tesztmágia) paradox módon a klasszikus tesztelmélet kidolgozottságának is következménye: a matematikailag kifinomult, a nem matematikus számára nehezen követhető, bonyolultságukkal tiszteletet parancsoló eljárások olykor feledtetik, elfedik az alapok esetleges ingatagságát.

Ha tehát a modern tesztelméletnek más haszna nem volna, mint az, hogy tudatosítja a klasszikus tesztelmélet alapján kidolgozott tesztek alkalmazásának *lehetőségei* mellett ezek *korlátait* is, már az is jelentős nyeresége volna a pszichológia tudományosságának. A modern tesztelmélet azonban a probabilisztikus modellek segítségével ennél többet szeretne elérni.

A látensstruktúra-modellek

A probabilisztikus modellek felé egy határozottan determinisztikus modell, L. Guttman perfekt skálája (perfect scale) tört utat: újszerűen kezelte a skálaproblémákat, a társadalomtudományokban szokásos adatok jellegének megfelelő matematikai módszereket keresett, és megfelelő matematikai modellel igyekezett megalapozni a skálaszerkesztési eljárást.

A Guttman-féle *perfekt skála* tulajdonságait jól mutatja egy-két egyszerű példa. A szociális státus (társadalmi-vagyoni helyzet) meghatározása történhet annak alapján, hogy kinek milyen használati tárgyak vannak a birtokában: van-e autója, színes televíziója, hűtőszekrénye stb. Hogyan lehet az ilyen jellegű (igennel-nemmel megválaszolható, de különféleképpen súlyozható) adatokkal skálát szerkeszteni? Akinek autója is van és színes televíziója is, az nyilván felsőbb rétegbe tartozik, magasabb helyet foglal el a szociális-vagyoni helyzet skáláján, mint akinek egyik sincs. Nem egyértelmű azonban a helyzet akkor, ha valakinek vagy az egyikre vagy a másikra tellett: ilyenkor egyéb tényezők is belejátszhatnak abba, hogy autót vagy színes tévét veszünk-e. Guttman hasonló esetekben úgy oldotta meg a problémát, hogy csakis olyan tárgyakat vett figyelembe, amelyek egyértelműen osztályoznak, elkülönítenek. Ha tapasztalati tény, hogy akinek autója van, annak hűtőszekrénye is van, akkor a közbülső fokozatot nem a színes televízió, hanem a hűtőszekrény birtoklásával kell mérni.

Hasonló elvek alapján vélemények, politikai állásfoglalások, attitűdök perfekt skálái is megkonstruálhatók. A szociális távolság mérésére szolgáló Bogardus-féle skála a legismertebb példa erre. Mindenki, aki igennel felel arra a kérdésre, hogy családtagként befogadna-e családjába egy négert (vagy más etnikai kisebbséghez tartozót), igennel fog válaszolni arra a kérdésre is, hogy munkahelyén együtt dolgozna-e egy négerrel (illetve a kérdéses etnikumhoz tartozó egyénnel). A fordítottja ennek nem igaz: a második kérdés igenléséből nem lehet az elsőnek igenlő megválaszolására következtetni. Így ez a két kérdés a perfekt skálán két különböző, egyértelműen meghatározott fokozatnak felel meg.

Guttman – mint látjuk – a szokott korrelációs módszerek helyett más utakat keresett a kvalitatív változók kezelésére. Ezzel annak idején (1941-ben) Paul LAZARSFELD (1968) szerint „Louis Guttman egy új világot tárt fel, hangsúlyozván annak szükségességét, hogy a kvalitatív adatokhoz új matematikai módszereket kell alkalmazni". A feltáruló „új világnak" fontos jellegzetessége volt a látens, külsőleg közvetlenül meg nem figyelhető dimenziónak (az intrinsic skálának) és a külsőleg megjelenő, különféle indikátorok felhasználásával megszerkeszthető (extrinsic) skálának megkülönböztetése. A mérés elméleti alapkérdése az extrinsic és intrinsic skáláknak *egymásra vonatkoztatása,* amely a kapcsolatukat közelebbről meghatározó *matematikai modell segítségével* történik.

Guttman elgondolása szerint az itemeket úgy kell súlyozni, hogy a klasszifikálandó objektumokat a lehető legélesebben elkülöníthessék. Ebből adódik a gyors népszerűségre szert tett modell eredendő gyöngéje: determinisztikus jellege. A skála egész szerkesztési eljárása és jogosultsága is arra a feltételezésre épült, hogy nincs kivétel: a magasabb értékű, nagyobb súlyú item igenlése, teljesedése automatikusan maga után vonja valamennyi alacsonyabb értékű, kisebb súlyú item teljesülését is. A kivételek eb-

ben a modellben nem erősítik, hanem megcáfolják a szabályt. Így azután viszonylag hamar kiderült, hogy Guttman modellje nem realisztikus, hiszen a társadalomtudományokban, a szociológiában, a pszichológiában figyelembe vehető mutatók (indikátorok) általában valószínűségi jellegűek: teljes biztonsággal sosem zárható ki, hogy például valaki, aki nagy autón járkál, hűtőszekrényt (mondjuk a frizsider-szocializmustól való félelmében) mégsem szerez be.

A *látensstruktúra-analízis* irányzata (mely főként Paul Lazarsfeld nevéhez kapcsolódik) ezért a valószínűségi, probabilisztikus szemlélet irányába fejlesztette tovább Guttman kezdeményezését.

A legegyszerűbb esetben ez a következőt jelenti. A különböző kérdésekre (például van-e autója? van-e színes tévéje? van-e hűtőszekrénye? stb.) adott igen-nem válaszok az egyénre jellemző válaszmintákat adnak, például ++; −+; +− stb. Az előforduló mintázatféleségek relatív gyakoriságából kiindulva a csoportosulások valószínűségére különféle számítások végezhetők. A cél a válaszolóknak látens osztályokba való besorolása. A besorolás nem determinisztikus: az egyénre vonatkozóan csak az a besorolási *valószínűség* (recruitment probability) határozható meg, amellyel valamely osztályba besorolódik. (Vagyis a besorolás valószínűségeloszlását határozzák meg különböző mintázatok esetén.) Az a priori nem ismeretes látens osztályokat a paraméterek egy halmaza jellemzi, amelyek minden egyes item és minden osztály vonatkozásában megadják a pozitív válasz valószínűségét arra az esetre, ha a válaszoló az illető osztályba beletartozik. (Az osztályozásnak ez a valószínűségi jellege különbözteti meg elvileg a látensstruktúra-analízist a cluster-analízistől.)

Minthogy a látens osztályok látens paraméterei előre nem ismeretesek, a megfigyelt adatokból kell következtetnünk rájuk, éspedig megfelelő matematikai modell felhasználásával. (A klasszikus tesztelmélet *közvetlenül* következtet a megfigyelt pontértékekből a „valódi értékekre" – a mérési hibahatárokon belül.) A látens paramétereknek és a megfigyelt válaszgyakoriságoknak összekapcsolódási módját a *modellegyenletek* (accounting equations) írják le. A modellegyenletek megoldásával a megfigyelt relatív gyakoriságokból becsülhetők az ismeretlen látens paraméterek.

Honnan nyerik a modellegyenleteket? Mire alapozódnak a látensstruktúrára vonatkozó matematikai modellek? Ezek bizonyos mértékig önkényesek: tartalmi és formai meggondolásokból eredő feltevésekre épülnek. Nagy előnyük a klasszikus tesztelmélettel szemben (mely ugyancsak meghatározott feltevésekre épül), hogy *ellenőrizhetők*. A látensstruktúra-analízis és más probabilisztikus modellek alkalmazásának elmaradhatatlan szakasza a modell érvényességének (az adatok illeszkedésének) vizsgálata, amelyre sokféle eljárást dolgoztak ki.

A probabilisztikus modellek előfeltevései közt *nem* szerepelnek a vizsgált látens tulajdonságnak a populációban való eloszlására vonatkozó feltevések, tehát nem szerepel a klasszikus tesztelmélet (ellenőrizhetetlen) normalitásfeltevése sem. Minden probabilisztikus modell szükséges előfeltevése viszont a „lokális sztochasztikus függetlenség".

Különböző valószínűségi változókat akkor mondunk egymástól függetlennek, ha az, hogy az egyik változó milyen értéket vesz föl, nem függ attól, hogy milyen értéket vett föl a többi. Vagy másként: több ($A_1, A_2, ..., A_n$) esemény együttes bekövetkezésének valószínűsége egyenlő az események külön-külön vett valószínűségei szorzatával

($P(A_1A_2...A_n) = P(A_1)\ P(A_2)..P\ (A_n)$). *Lokális* függetlenségen pedig az értendő, hogy egy látens osztályon belül fennáll a sztochasztikus függetlenség, vagyis a látens osztályba tartozó részpopulációban az egyik itemre adott pozitív, illetve negatív válasz nem függ a többi itemre adott válaszoktól. Jelöljük az *i*-edik kérdésre adott pozitív, illetve negatív válasz valószínűségét a következőképpen: $p\{(\overset{i}{+})\}, p\{(\overset{i}{-})\}$. Ekkor a következőképp írható fel annak valószínűsége, hogy a látens osztályba beletartozó személy az *i*-edik és a *j*-edik feladatra miként válaszol:

$$p\{(\overset{i}{+}), (\overset{j}{+})\} = p\{(\overset{i}{+})\}\ p\{(\overset{j}{+})\}$$

$$p\{(\overset{i}{+}), (\overset{j}{-})\} = p\{(\overset{i}{+})\}\ p\{(\overset{j}{-})\} \text{ stb.}$$

Ez azt is jelenti, hogy az itemek közt — egy látens osztályon belül — nincs korrelációs kapcsolat. (A lokális sztochasztikus függetlenség feltevésére azért van szükség, mert egy megfigyelt esemény relatív gyakorisága csak abban az esetben szolgálhat az események valószínűségének becsléséül, ha a megfigyelések függetlenek, egymást nem befolyásolják.)

A látens osztályok megkeresésének szempontja ennek megfelelően éppen az, hogy az itemek közt a teljes sokaságban még meglevő korrelációt (asszociációt) az osztályokra bontással megszüntessék (mintegy „kiparcializálják"). Ami azzal egyértelmű, hogy a korrelációt bizonyos látens osztályokhoz való tartozásra (látens dimenzió mentén való elhelyezkedésre) vezették vissza: a teljes sokaságban az itemek közt megfigyelt statisztikus összefüggés egy közös látens változótól való függés eredményének tekinthető. (A.K. Formann előadásaiból vett példával: a lokális sztochasztikus függetlenség szemléletmódja leginkább azzal illusztrálható, hogy a betegségtünetek a betegség révén függnek csak össze egymással, noha az orvos szemében egybetartozó tünetegyüttest alkotnak.)

Lássunk egy fiktív számpéldát is. Olvasási szokások vizsgálatából kiderül, hogy a megkérdezett 216 pedagógus közül 54-en a Pedagógiai Szemlét is és a Pszichológiai Szemlét is, 102-en egyiket sem, 30—30-an pedig vagy egyiket, vagy másikat olvasgatják. Következik-e a pozitív korrelációból, hogy a pedagógiai és pszichológiai érdeklődés általában együtt jár? Ha a „továbbképzésben való részvétel" látens dimenziója alapján két osztályra bontjuk a megkérdezetteket, az így nyert két osztályban a korreláció eltűnt. A korrelációt a látens dimenzióra vezettük vissza: a pedagógiai és pszichológiai érdeklődés általában nem jár együtt, a tapasztalható együttjárás a továbbképző tanfolyamon való részvétel hatása *(4. táblázat)*.

Példáink eddig látens osztályokra vonatkoztak. Tesztek esetében azonban nem diszkrét, hanem folytonos látens dimenzióra, egy látens kontinuumra kell gondolnunk. A lokális sztochasztikus függetlenség feltevése ilyenkor is hasonlóan értelmezendő: a kontinuum azonos pontján elhelyezkedő (ott lokalizálható) személyek összességére áll, hogy az itemekre adott válaszaik függetlenek egymástól. A tesztitemek közt a teljes populáció figyelembevételekor tapasztalható korreláció a látens dimenzióra (képességre, intelligenciára) vezethető vissza. Ha a látens paraméter változatlan (rögzítve van, mert a kontinuumnak csak egy pontját tekintettük), a korreláció eltűnik.

Egyes probabilisztikus modellek (így a Rasch-modell is) további kiegészítő fel-

4. táblázat

Négyzetes kontingencia-táblázatok a Pedagógiai Szemle és a Pszichológiai Szemle olvasóiról. a) A teljes populációra vonatkozóan; b) a továbbképzés résztvevőire; c) a továbbképzésen részt nem vevőkre vonatkozóan

Pedagógiai Szemle

Pszichológiai Szemle

	+	–
+	54	30
–	30	102

a) $r > 0$

	+	–
+	50	10
–	10	2

b) $r = 0$

	+	–
+	4	20
–	20	100

c) $r = 0$

tevéssel élnek: eszerint a megoldott tesztitemek száma a vizsgált személyek jellemzése szempontjából „elégséges statisztikát" (sufficient statistic, erschöpfende Statistik) alkot. A R. Fishertől származó fogalom értelme az, hogy — a feltevés szerint az adott modellben — nem okoz információveszteséget, ha valamennyi itemmegoldást egységnyi súllyal vesszük számításba. Vagyis például egyenértékűnek (=2) vesszük a +–+; ++–; –++ megoldássorokat. (Ez a feltevés nemcsak a probabilisztikus tesztmodellekben fordul elő, hanem számos más, hagyományos teszt szerkesztésekor is hallgatólagosan élnek vele.) Bár minden helyes feladatmegoldás minden item esetében (az „elégséges statisztika" elvének megfelelően) egyformán 1 pontértékkel járul hozzá a személy teljesítményét jellemző összpontszámhoz, ebből általában nem következik, hogy a feladatoknak azonos nehézségűeknek kell lenniük. (Olyasmi ez, mint az olimpiákon a nemzetek rangsora: csak az aranyérmek számítanak, függetlenül attól, hogy atlétikában, futballban vagy íjászatban szerezték-e őket. Ahogy a sakkversenyen is csak a szerzett pontok száma dönt, függetlenül attól, hogy az erősebb játékosokat győztük-e le, s a gyengébbektől kaptunk-e ki vagy fordítva. A tenisz ranglistái viszont nem ilyen egyszerűsítő meggondolások szerint készülnek.)

A probabilisztikus tesztmodellek fontos jellemzői az item-*jellegfüggvények*. A modellek probabilisztikus szemléletmódja mutatkozik meg abban, hogy a jellegfüggvények valószínűség-eloszlásfüggvények: azt írják le, hogy valamely adott (tehát rögzített nehézségi szintű) *i* feladat megoldásának *valószínűsége* miként változik a feladatmegoldó személy „képességének" (ξ) függvényében:

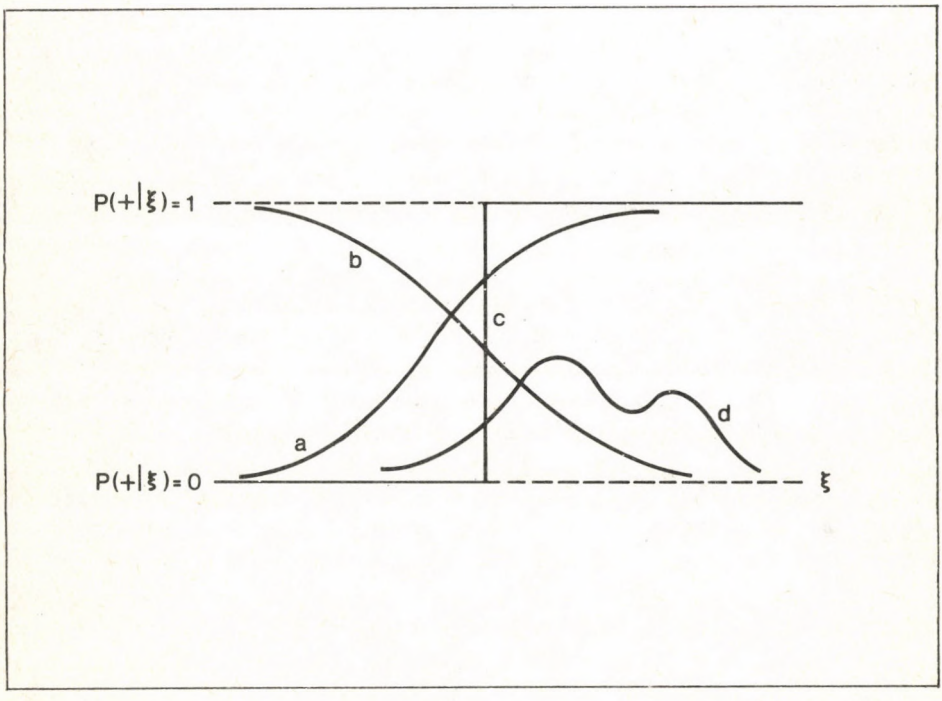
$$p\left\{(i_+) \mid \xi\right\} = f_i(\xi)$$

Mit fejez ki tehát egy item jelleggörbéje? Mutatja annak valószínűségét, hogy a ξ látens „képesség"-kontinuum bizonyos (például ξ_1 vagy ξ_2 pontján) elhelyezkedő személy milyen valószínűséggel oldja meg helyesen az *i* feladatot. Valószínűség-eloszlásfüggvény lévén a jelleggörbe monoton növekszik (tehát $\xi_1 < \xi_2$, akkor $P\left\{(i_+) \mid \xi_1\right\} <$

$< P\left\{ (^{i}_{+}) \mid \xi_2 \right\}$ és aszimptotikusan 0-hoz, illetve 1-hez tart (ha ξ a $-\infty$-hez, illetve a $+\infty$-hez közeledik).

Az egyes feladatokhoz tartozó eloszlásfüggvények közelebbi meghatározása (specifikálása) a modell leírásának fontos mozzanata. Az így kapott ún. *jelleggörbék* (traceline, operational characteristic, Item-charakteristik) egy modellen belül sem feltétlenül azonos típusúak (bár többnyire azok), a különböző tesztmodellek pedig főként az itemek jellegfüggvényeinek (jelleggörbéinek) megválasztásában különböznek egymástól.

Guttman determinisztikus modelljében a jelleggörbék nem ilyen tulajdonságúak (lásd *2. ábra*); vagy 0 vagy 1 értéket vesznek fel és egy meghatározott pontban szaka-



2. ábra
Két szigorúan monoton item-jelleggörbe (*a* és *b*), egy Guttman-item jelleggörbéje (*c*), és egy nem monoton item (*d*) jelleggörbéje (FISCHER, 1974, 155. o.)

dásuk van. Tukey és Lord modelljében a jelleggörbe ún. normál ogiva, Birnbaumnál és (részben Raschnál is) pedig ún. logisztikus görbe. Birnbaumtól eltérően Rasch feltételezi, hogy valamennyi item jelleggörbéje párhuzamos futású; Rasch modellje felfogható a Birnbaum-modell sajátos eseteként is.

Az, hogy a jelleggörbe alakjára vonatkozó feltevések megfelelnek-e az adatok jellegének, természetesen nem dönthető el eleve: ennek ellenőrzésére a modelltesztek hivatottak.

Fontos, hogy ne tévesszük össze az itemek megoldási valószínűségére vonatkozó eloszlásfüggvényt a vizsgált populáció képességeinek (például a normális eloszlást közelítő) eloszlására vonatkozó feltevéssel. Ezt a tévedést követi el GUTJAHR (1974, 240. o.), amikor egy ún. Rasch-eloszlást posztulál és megállapítja, hogy az nem tér el lényegesen a normális eloszlástól. Valójában az item-jelleggörbének a populáció jellegéhez semmi köze, és a Rasch-modell alapvető tulajdonsága az, hogy benne az itemparaméterek a populáció személyparamétereitől éppúgy függetlenül becsülhetők, mint ahogy a személyparaméterek is függetlenek az itemparaméterektől.

A Rasch-modell

Könyvében RASCH (1960) három különböző modellt írt le. Az egyik az olvasási hibák felmérésére, a másik az olvasás sebességének vizsgálatára, a harmadik pedig intelligenciatesztekre vonatkozott. Mindhárom modell probabilisztikus jellegű volt; a megfigyelt esemény – a tévesztések, az adott idő alatt olvasott szavak, a helyes tesztfeladatmegoldások – bekövetkezésének valószínűségére egy-egy eloszlást adott meg egy-egy paraméter segítségével (amely paramétert mindhárom modell két egymástól függetlenül változó tényező szorzataként határozott meg). A három modell az itemkarakterisztika-függvény (a jelleggörbe) megválasztásában tért el egymástól: az első két modellben a Poisson-eloszlás, a *harmadik* modellben egy formailag igen egyszerű függvény szerepelt. E függvény megválasztása „szerencsésnek" bizonyult: a későbbiekben bebizonyosodott, hogy fontos modelltulajdonságok következnek belőle. Ma Rasch-modell néven (különböző, de matematikailag ekvivalens megfogalmazásokban) ezt az eredetileg is intelligenciavizsgálatok (a dán hadseregben használt tesztek) anyagán kidolgozott modellt emlegeti a szakirodalom.

Ez a Rasch-modell a következő három *előfeltevésen* alapul:

„1. Az *i*-edik tesztfeladatra válaszoló *v* személyhez minden tesztszituációban a helyes válasz meghatározott valószínűsége tartozik, amelyet az alábbi formában írhatunk fel:

$$p\,(+\mid v, i)=\frac{\lambda_{vi}}{1+\lambda_{vi}}, \quad \lambda_{vi}\geqslant 0.$$

2. A λ_{vi} állapotparaméter két mennyiség szorzata

$$\lambda_{vi}=\xi_v\ \epsilon_i,$$

ahol ξ_v a személyre, ϵ_i pedig a feladatra vonatkozik.

3. Adott rögzített paraméterek mellett a válaszok sztochasztikusan függetlenek egymástól" (RASCH, 1973, 92. o.).

Nézzük meg, hogyan változik a feladatmegoldás valószínűsége

$$f_i(\xi_v) = p(+ \mid v, i) = \frac{\xi_v \epsilon_i}{1 + \xi_v \epsilon_i} = \frac{\frac{\xi_v}{\delta_i}}{1 + \frac{\xi_v}{\delta_i}} = \frac{\xi_v}{\xi_v + \delta_i}$$

a személyparaméter, illetve a feladatparaméter ($\epsilon_i = \frac{1}{\delta_i}$) változásával. (A feladatparamétert reciprok alakban is megadhatjuk; az 1960-as leírásban így szerepelt.) Az $f_i(\xi_v)$ függvény csak 0 és 1 közötti értéket vehet föl; ha a ξ_v személyparaméter 0-hoz közeli értéket vesz föl, amit úgy értelmezhetünk, hogy rendkívül gyönge „képességű" személlyel van dolgunk, a helyes megoldás valószínűsége ugyancsak zéróhoz közeli mennyiség. Ha δ_i – amit a feladat „nehézségének" tekinthetünk (s ennek megfelelően reciproka, ϵ_i az i-edik feladat „könnyűségének" mutatója) – a 0 felé tart, a helyes feladatmegoldás valószínűsége egyhez közeledik. 1-hez közeledik a megoldási valószínűség akkor is, ha a „képesség"-paraméter minden határon túl növekszik. A λ_{vi} paraméter zérótól a végtelenig minden értéket felvehet, s így igen könnyű és igen nehéz problémák, valamint szerény képességűek és kiválóak megoldási valószínűségei egyaránt jellemezhetők.

Rasch külön hangsúlyozza, hogy a „képesség"-(ξ) és a „nehézség"-paraméter (δ) csak együttesen vezethető be modelljében: úgy tartoznak össze, oly módon szimmetrikusak, mint Newton második törvényében a tömeg és az erő fogalma: $\lambda = \frac{\xi}{\delta}$, miként $a = \frac{p}{m}$ (a-val a gyorsulást, p-vel az erőt, m-mel a tömeget jelölve). A Newton-törvénnyel analóg módon „multiplikatív törvény" formájában írhatók föl (mindhárom) Rasch-modellben az alapösszefüggések, így elvben a tömeg és erő méréséhez analóg módon (hasonló fogalmi struktúrával) a modellparaméterek méréséhez is arányskála konstruálható. Ennek következtében – bár a fizikai törvény determinisztikus, a tesztmodell probabilisztikus jellegű – „egy gyermek olvasáskészsége ugyanolyan objektivitással mérhető, mint a testsúlya, legfeljebb nem olyan precizitással, dehát ez már más kérdés" (RASCH, 1960, 115. o.).

Ha $\xi'_v \epsilon'_i = \xi_v \epsilon_i$ ($= \lambda_{vi}$), akkor a $\frac{\xi'_v}{\xi_v} = \frac{\epsilon_i}{\epsilon'_i}$ arány állandó a v és i bármely kombinációja esetén. Ha valamely feladat (például $i = 0$) nehézségét mércéül választjuk („egységnyi" nehézségűnek), ezzel a személyparaméter egységét is megválasztottuk, tehát elvileg ugyanúgy járunk el, mint a fizika a tömeg és az erő esetében.

A tényleges „mérés" azonban csak akkor oldható meg, ha a kétféle paraméter bizonyos értelemben *elkülöníthető* egymástól. Vagyis bizonyos számú személynek bizonyos feladatokra adott válaszaiból a mennyiségek két különböző halmaza származtatható; az egyiknek az eloszlása csakis a feladatok paramétereitől függ, és nem függ a személyparaméterektől, a másik pedig csakis a személyek képességparamétereitől függ és nem függ a feladatparaméterektől.

Vegyünk egy olyan tesztet, mely két itemből (i és j) áll. Egy ξ paraméterű sze-

mély esetében (a sztochasztikus függetlenség feltevésének megfelelően) annak valószínűsége, hogy mindkét feladatot helyesen (+,+) megoldja, a megfelelő valószínűségek szorzatával egyenlő.

$$p\left\{(\overset{i}{+}), (\overset{j}{+}) \mid \xi\right\} = p\left\{(\overset{i}{+}) \mid \xi\right\} \cdot p\left\{(\overset{j}{+}) \mid \xi\right\} = \frac{\xi^2 \epsilon_i \epsilon_j}{(1+\xi\epsilon_i)(1+\xi\epsilon_j)} \qquad (I)$$

Hasonlóképpen

$$p\left\{(\overset{i}{+}), (\overset{j}{-}) \mid \xi\right\} = \frac{\xi \epsilon_i}{(1+\xi\epsilon_i)(1+\xi\epsilon_j)} \qquad (II)$$

$$p\left\{(\overset{i}{-}), (\overset{j}{+}) \mid \xi\right\} = \frac{\xi \epsilon_i}{(1+\xi\epsilon_i)(1+\xi\epsilon_j)} \qquad (III)$$

$$p\left\{(\overset{i}{-}), (\overset{j}{-}) \mid \xi\right\} = \frac{1}{(1+\xi\epsilon_i)(1+\xi\epsilon_j)} \qquad (IV)$$

Annak a valószínűségét, hogy mindkét feladatot helyesen oldja meg a tekintett személy, az első formula adja meg, annak valószínűségét, hogy egyiket sem oldja meg, a negyedik formulából olvashatjuk le. A második és harmadik formula *összege* viszont azt a valószínűséget adja meg, hogy kettő közül valamelyik feladatot helyesen, a másikat helytelenül oldja meg:

$$P\left\{\text{csak az egyik megoldása} + \mid \xi\right\} = \frac{\xi(\epsilon_i + \epsilon_j)}{(1+\xi\epsilon_i)(1+\xi\epsilon_j)} \qquad (V)$$

A feltételes valószínűségekre vonatkozó összefüggések ($P(A|B)=P(AB)|P(B)$) alapján a (II) és (V) hányadosaként kaphatjuk meg annak valószínűségét, hogy épp az *i*-edik feladat megoldása helyes, feltéve, hogy a ξ képességű személy valamelyik feladatot helyesen oldja meg, a másikat viszont nem:

$$p\left\{(\overset{i}{+}) \mid \pm, \xi\right\} = \frac{\xi \epsilon_i}{(1+\xi\epsilon_i)(1+\xi\epsilon_j)} \cdot \frac{(1+\xi\epsilon_i)(1+\xi\epsilon_j)}{\xi(\epsilon_i+\epsilon_j)} = \frac{\epsilon_i}{\epsilon_i+\epsilon_j}$$

Vagyis az egyenlet jobb oldaláról a ξ paraméter kiesett: a vizsgált valószínűség *nem függ* a személyparamétertől, hanem *csakis* a két feladatparamétertől. Teljesen szimmetrikus formula vezethető le a személyparaméterek vonatkozásában is: a

$$p\left\{(\overset{i}{+}) \mid \pm, \epsilon\right\} = \frac{\xi_w}{\xi_w + \xi_v}$$

valószínűség nem függ attól, hogy melyik (milyen nehézségű) feladat felhasználásával vizsgálják.

A modellnek ez a sajátossága általánosítható tetszőleges *k* számú kérdésre és *n* számú személyre. Ehhez kihasználjuk az „elégséges statisztika" (sufficient statistic, erschöpfende Statistik) feltevését, amely szerint a pontértékek az adathalmaz által tartalmazott minden információt magukban foglalnak (a +-ok és – -ok váltakozásából adódó mintázatok a + számához képest nem adnak további információt). Miként azt FISCHER (1974, 193. o.) megmutatta, az elégséges statisztika a Rasch-modell szükséges és elegendő feltétele (tehát egyrészt az alapfeltevésekből levezethető, másrészt a modell másik két alapfeltevésének teljesülése esetén szükségképpen a Rasch által feltételezett eloszlásfüggvényhez vezet).

A Rasch-modellben szereplő $f_i(\xi)$ függvényt szokás más (nem multiplikatív, hanem additív), de matematikailag ekvivalens alakban is felírni (amelyből azonban nem arányskála, hanem differenciaskála vezethető le):

$$f_i(\xi) = \frac{\exp(\xi - \delta)}{1 + \exp(\xi - \delta)}.$$

Ez a formula egyszerűen adódik az eredetiből

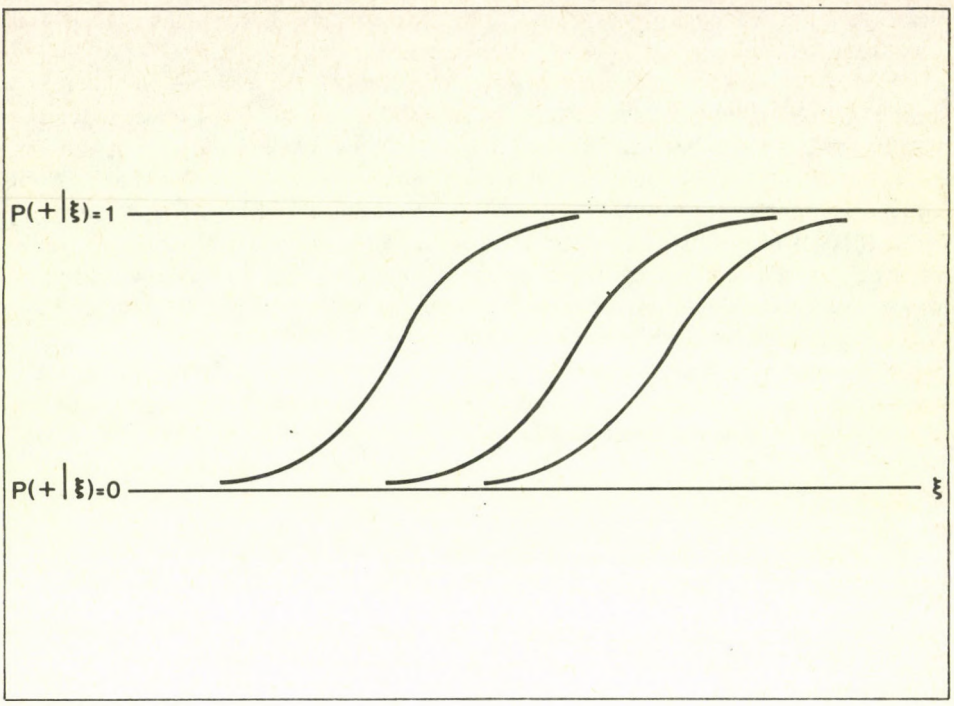
$$f_i(\xi) = \frac{\lambda}{1+\lambda} = \frac{\xi\epsilon}{1+\xi\epsilon} = \frac{\frac{\xi}{\delta}}{1+\frac{\xi}{\delta}} = \frac{e^{\ln\xi - \ln\delta}}{1+e^{\ln\xi - \ln\delta}} = \frac{\exp(\xi-\delta)}{1+\exp(\xi-\delta)}.$$

A Rasch-modellt ebben az alakjában „speciális logisztikus modellnek" is szokták nevezni. (Megkülönböztetésül Birnbaum ún. általános logisztikus modelljétől, amelyben nem két, hanem három paraméter szerepel, s amely Rasch-modelljénél bonyolultabb. A Rasch-modell a Birnbaum-féle modell speciális esetének tekinthető.)

Az $f_i(\xi)$ függvény megadja a ξ kontinuum bármely kiválasztott pontjához az *i*-edik feladat megoldásának valószínűségeloszlását, amelyet logisztikus görbe ábrázol. A különböző feladatok jelleggörbéi – a Birnbaum-modelltől eltérően – a Rasch-modellben párhuzamosak *(3. ábra)*.

Az $f_i(\xi)$ eloszlásfüggvény (jelleggörbe) ismeretében tesztfeladatok megoldására vonatkozó adatokból maximum-likelihood eljárással *becsülhetők* az itemparaméterek. (A maximum-likelihood becslés lényegéről lásd HAJTMAN, 1968, 472. o.) A személyparaméterek analóg módon becsülhetők volnának az alapadatokból, általában azonban az itemparaméterek ismeretében – a teszthez kidolgozott táblázat felhasználásával – határozzák meg a személyparaméterek becsült értékeit.

Az additív formában felírt modellnél – mint láttuk – a feladatmegoldás elméleti valószínűsége a két paraméter különbségétől függ, nem pedig abszolút értéküktől: csak két paraméter differenciája van egyértelműen meghatározva. A paraméterskála nullpontját tetszőlegesen választhatjuk meg. Gyakran úgy normálják a tesztet, hogy az itemparaméterek összege zéró legyen. Ekkor a negatív itemparaméterek könnyű itemeket, a pozitív értékek nehéz feladatokat jelölnek (amikor is az itemek többsége egy –3 és +3 közötti skálán helyezhető el). A személyparamétereknél analóg a helyzet: pozitív



3. ábra
Rasch speciális logisztikus tesztmodelljének jelleggörbéi, melyek eltolással fedésbe hozhatók
(FISCHER, 1974, 199. o.)

paraméterértékűek a vizsgált feladattípus vonatkozásában „jó képességű" személyek. (A modell multiplikatív alakja arányskálához vezet, amelynek normálása is más.) Egy személyparaméter annál pontosabban becsülhető meg, minél nagyobb számúak az itemek, és nehézségi szintjük minél inkább megfelel a személy képességszintjének. A személyparaméter becslésének pontossága (konfidenciaintervalluma) tehát attól is függ, hogy az a „képesség"-skála mely tartományában helyezkedik el. (A klasszikus tesztelmélet – a tapasztalati tények ellenére – minden tulajdonságfokozat esetén, vagyis a skála minden pontjánál ugyanazt a mérési hibát feltételezi, amelynek nagysága a tulajdonság teljes mintabeli varianciájától függ.)

A Rasch-modell két paramétertípusának szeparálhatóságából (tehát a személyparamétereknek a feladatparaméterektől és a feladatparamétereknek a személyparaméterektől való függetleníthetőségéből) következik, hogy a Rasch-modell feltételeinek megfelelő tesztek eleget tesznek a „specifikus objektivitás" feltételének (ami a klasszikus tesztekről általában nem mondható el). A *specifikus objektivitás* azt jelenti, hogy ha két személyt összehasonlítunk, akkor az összehasonlítás eredményének nem szabad függenie attól, mely (homogén) tesztitemek alapján történik az összehasonlítás. Hasonlóképp a tesztitemek paramétereinek meghatározásakor sem függhet az itemek összehasonlítása attól, hogy milyen képességű személyek közreműködésével

történik ez. Vagyis az itemparaméterek függetlenek a mintapopuláció képességeinek eloszlásától és viszont.

A modellnek ezek a tulajdonságai teszik a modellt *ellenőrizhetővé,* ami a Rasch-modell egyik lényeges előnye (a klasszikus tesztelmélet implicit vagy explicit feltevéseinek ellenőrizhetetlenségével szemben). Számos különböző *modellpróbát* dolgoztak ki, ezek közül a legfontosabbak a likelihoodhányados-tesztek. Az ellenőrizendő modellnél meghatározzák az adatokhoz tartozó „valószerűség"-értéket (likelihoodot)* olyan paraméterbecslések alapulvételével, amelyek a „valószerűséget" maximálják (tehát a maximális likelihoodot számolják ki). Ezután a modell egyes korlátozó, megszorító feltételeit elhagyják. Kevesebb megszorítást tartalmazó modell esetében az adatokhoz tartozó „valószerűség"-érték (likelihood) — amelyet az előzőkhöz hasonló módon ismét kiszámolnak — csak nagyobb lehet, mint az előbb volt. A két likelihood — az első és a második — hányadosa így 0 és 1 között változhat, és értéke annál kisebb, minél kevésbé feleltek meg az elhagyott modellfeltevések az adatoknak, vagyis minél erősebben növekedett a módosítás következtében az adatokhoz tartozó likelihood. (A modelleltérések statisztikai szignifikanciája a szokásos módon vizsgálható.)

A Rasch-modell likelihoodhányados-módszerrel való ellenőrzésekor többnyire azt a feltevést hagyják el, hogy az itemparaméterek a személyrészmintákban ugyanolyan nagyok, hiszen ez a feltevés, a populációfüggetlenség a modell alapvető követelménye. Ennek a követelménynek teljesülése ellenőrizhető (legalább hozzávetőlegesen) egyszerűbb, grafikus módszerrel is, amit már Rasch is alkalmazott. Ha a populációt bármely szempont szerint (például gyengébbek és jobb képességűek) két részre osztjuk, és a két részre vonatkoztatva újból meghatározzuk az itemparamétereket, azok lényegesen nem különbözhetnek egymástól. Tehát ha a koordináta-rendszer egyik tengelyére az egyik csoportban nyert, a másik tengelyre a másik csoportban nyert itemparamétereket tüntetjük fel, az összetartozó koordinátájú pontok hozzávetőlegesen egy egyenes mentén helyezkednek el.

A modell érvényessége azt is jelenti, hogy az itemek *homogének,* ugyanazt a képességet mérik. Ha a modellpróbából előbb az derül ki, hogy a modell (még) nem érvényes, akkor a gyanús itemek elhagyásával esetleg a homogenitás mégis elérhető. A modelleltérések okainak elemzése ugyancsak értékes tudományos felismerésekkel járhat, ezért a Rasch-modell, illetve a modellpróba abban az esetben is hasznos kutatási eszköznek bizonyulhat, amikor valamilyen okból nem sikerült az itemek homogenitását, a paramétereknek a „mintától" való függetlenségét biztosítani. Mindenesetre — Rasch szavaival — „meglepően sokszor" derült ki konkrét vizsgálatokban a Rasch-modell érvényessége.

*A likelihood szótári jelentése magyarul — éppúgy mint a probability-é — 'valószínűség'. A magyar nyelvű szakirodalomban azonban gyakran fordítják „valószerűségnek", hogy, akárcsak R. Fischer tette, megkülönböztessék a probability-vel jelölt valószínűségtől. Mert bár a likelihood is bizonyos értelemben valószínűséget fejez ki, a likelihood-függvény *nem* valószínűségeloszlás-függvény, matematikai tulajdonságai mások.

A binomiális modell

A binomiális modell alapján szerkesztett teszteket elsősorban az ún. „mastery learning" (zielerreichendes Lernen; eredményes tanulás, vagy Nagy József magyarításában: megtanítási stratégia) igényei hívták életre. A modell kidolgozója K.J. Klauer, a „mastery learning" céljaira alkalmazta többek közt EIGLER és STRAKA (1978).

A „mastery learning" célkitűzése, hogy alapvető ismereteket (képességeket) minden egyes tanulóval elsajátíttasson (minden tanulóban kifejlesszen). A tesztelés célja itt tehát nem az, hogy a tanulókat valamiképp rangsorolja (amit a klasszikus „reálnormaorientált" tesztek célul tűznek ki), hanem hogy megvizsgálják, elsajátította-e a tanuló a mindenki számára szükséges ismereteket („eredményorientált" tesztelés).

Az elsajátítás ismérve, hogy a tanuló nagy valószínűséggel meg tud oldani bizonyos feladattípushoz tartozó feladatot. Mekkora valószínűséggel? Minthogy az 1 valószínűség követelménye nem volna reális, ennél szerényebb követelménnyel kell beérni. Klauer $p_z = 0{,}9$ vagy 0,95 valószínűséget javasol követelményként. A teszttel azt kell felbecsülni, hogy valamely (i-edik) tanulónál az adott feladattípushoz tartozó feladatok megoldásakor a helyes megoldás valószínűsége (p_i; az i-edik tanuló kompetenciáját jellemző érték), eléri-e a célul kitűzött p_z értéket. Ha nem, akkor további felzárkóztató tanulásra van szükség.

Mivel egyetlen feladat helyes vagy helytelen megoldása alapján a valószínűséget becsülni nem tudjuk, több azonos típushoz tartozó és *azonos nehézségi szintű* feladat megoldatására van szükség; minél több feladatot alkalmazunk, annál pontosabb lehet a becslés. A becslés hibája (egy megadott α tévedési valószínűséggel) szintén megbecsülhető, tehát meghatározható a konfidenciaintervallum. Ehhez a valószínűségek binomiális eloszlása használható fel; ez indokolja *a binomiális modell* elnevezést.

Egy valószínűségi változót akkor nevezünk n-edrendű, p paraméterű binomiális eloszlásúnak, ha lehetséges értékei a 0, 1, 2, ..., n egész számok, és éppen a k értéket (amely tehát 0 vagy 1 vagy 2 ... vagy n lehet)

$$B_{n,k} = \binom{n}{k} p^k (1-p)^{n-k}$$

valószínűséggel veszi fel. Ha n tesztfeladat közül az i-edik tanuló épp k feladatot oldott meg helyesen, p ismeretében a binomiális eloszlás alapján meghatározható, menynyi annak a valószínűsége, hogy tanulónk véletlenül jutott ehhez az eredményhez, tehát n-ből k helyes válaszhoz. A tesztelés célkitűzésével összhangban azt kell meghatározni, hogy *legalább hány* feladat megoldása esetén tekinthető a kapott eredmény (kellő valószínűséggel) *nem* véletlenül nyert, hanem a megtanulás nyomán bekövetkezett eredménynek.

A statisztikai hipotézisvizsgálatok szokott gondolatmenetével és a binomiális eloszlás táblázatainak felhasználásával egyszerűen meghatározható (adott tévedési valószínűséggel), hogy legalább hány helyesen megoldott feladat esetén lehet a tanulási célt elértnek tekinteni.

Mint láttuk, a binomiális modell előfeltevései, illetve a tesztfeladatokkal szembeni követelményei a következők:

1. A feladatmegoldások itt is csak „helyes – nem helyes'' alternatívával értékelhetők.

2. A feladatsokaság, amelyből a teszteléshez válogathatunk, (elvileg) végtelen nagy. (Gyakorlatban feladattípusonként legalább 60 feladatból álló feladatválasztékot szoktak megkívánni.)

3. A feladatmegoldások egymástól függetlenek legyenek (sztochasztikus függetlenség), azaz az egyik feladat megoldása ne befolyásolja a másik feladat megoldásának valószínűségét.

4. A teszt feladatai ugyanahhoz a feladattípushoz tartozzanak, vagyis ugyanazt a célul kitűzött tanulási eredményt vizsgálják (homogenitás).

5. A vizsgált személy számára bármely tesztfeladat megoldási valószínűsége ugyanaz legyen; tehát a feladatok azonos nehézségűek legyenek.

Ezek az előfeltevések egy lényeges pontban különböznek a Rasch-modelltől. Ez az utolsó pont, az azonos nehézségi szint követelménye. A binomiális modellben a feladatok *jelleggörbéi fedik egymást.* A binomiális modell a Rasch-modell speciális esetének is tekinthető tehát (ami azt is jelenti, hogy a Rasch-modellhez alkalmazott modellpróbák itt is felhasználhatók). A modell speciális (5.) feltevésének teljesülése ugyancsak könnyen ellenőrizhető: a helyes megoldást elérők száma (véletlen ingadozásoktól eltekintve) minden feladatnál ugyanannyi kell hogy legyen.

Alkalmazások és távlatok

Az az ismertetés, amelyet a probabilisztikus tesztmodellekről adhattunk, minden szempontból vázlatos. Az elmúlt tíz-tizenöt esztendőben a Rasch-modellt is több irányban továbbfejlesztették, illetve kiterjesztették. E fejleményekből sok mindent tartalmaz már FISCHER (1974) vaskos monográfiája is, de az azóta eltelt évtized az újabb tesztmodellek lehetőségeinek további kipróbálását is eredményezte. A fejlődés egyrészt a matematikai megalapozásra, a szűkebb értelemben vett matematikai statisztikai, számítástechnikai eljárásokra (a paraméterek becslésére, a modellkonformitás próbáira) terjedt ki. Másrészt, az eredetileg dichotom (igen-nem) válaszokkal és két (személy-, illetve item-) paraméterrel számoló modellnek Fischer kidolgozta *polichotom* (többfokozatú válaszokat is megengedő) változatát is, és megmutatta, miként lehet az eredeti két paraméteren túl – bizonyos további megkötések révén – újabb paramétereket is a vizsgálatokba bevonni *(lineáris logisztikus tesztmodell).*

Nagyszámúak a közvetlen alkalmazási kísérletek is. Kiterjedt, intenzív munkálatok folytak és folynak a Rasch-modell alapján készült *tantárgytesztekkel*, vizsgaanyagokkal. A Chicagoi Egyetemen Ben Wright és munkatársai, Angliában a National Foundation for Educational Research keretében kutató WILLMOTT és FOWLES (1974) tartozik az úttörők közé. (Az NSZK-ban, mint már említettük, Klauer, Straka és mások a számítástechnikai szempontból is egyszerűbb binomiális modellel dolgoznak.)

A Rasch-modell hasznosnak bizonyult a pszichológiai kérdőívek elemzésében, *attitűdvizsgálati eszközök* kidolgozására is. (A nagyszámú ilyen munka közül kiemelkedik WAKENHUT (1974) monográfiája, amely részletesen leírja az attitűdskála szer-

kesztésének egész menetét, és egybeveti a Rasch-modellen alapuló eljárást a Likert-skálával.)

Új lehetőségeket jelent a Rasch-modell alkalmazása a változásmérésekben, miként azt Ilse Rop vizsgálatai szemléltetik. (Figyelemre méltó „melléktermékei" például Rop vizsgálatainak azok az adatok, amelyek a Piaget-féle fejlődési fázisokkal kapcsolatos vitában Bruner álláspontját erősítik.) Új módon közelíthetők meg a Rasch-modell alapján az olyan sokat vitatott kérdések is, mint az intelligencia differenciálódási hipotézise (Spada adatai szerint ez a differenciálódás 14 éves életkor után – az általa alkalmazott teszt vonatkozásában – csak csekély mértékű).

Fischer, Spada és mások *projektív tesztek* (Rorschach, Holtzman, TAT) vizsgálatára is alkalmazták a probabilisztikus modelleket, hogy ezáltal a projektív teszteket „a hit vagy nem hit tartományából az egzakt tudományosság tartományába vonják". (A többi között megállapították, hogy a személy- és táblaparaméterek az értelmezési folyamatban ugyanúgy együttesen hatnak, miként a „képesség" és a „nehézség" az intelligenciatesztek feladataiban. A Rorschach-teoretikusoknak az értelmezési folyamat komplexitásáról alkotott „viszonylag amorf elképzeléseivel" szemben tehát a projektív válaszadást egy „viszonylag egyszerű törvényszerűségre" vezethették vissza.)

A *lineáris logisztikus tesztmodell* jól használható feladat*struktúrák* (például tesztitemek) *bonyolultságának* mérésére, és (a megfelelő kikötések teljesülése esetén) alkalmas eszköz lehet gondolkodáspszichológiai és tanulási kísérletek nehézségi szintjének megtervezésére, illetve elemzésére is, amire érdekes példa KUBINGER (1980) tanulmánya.

A Rasch-modell tehát valóban meglepően sok területen „működik". Hasznos vizsgálati eszköz lehet azonban ott s akkor is, ahol „nem működik". Amikor egy itemkészletből *a modellpróbák* nyomán itemeket modellösszeférhetetlenség miatt ki kell selejtezni, a kiszűrt itemek olyan tulajdonságait tárhatja fel az elemzés, amelyre a modellvizsgálat nélkül nem derült volna fény. (Ugyanez vonatkozhat a „személyösszeférhetetlenség" eseteire is.)

A probabilisztikus modellek fejlődése eddig is sokat köszönhetett a számítástechnikai forradalomnak; enélkül a gyakorlati alkalmazás reménytelen lett volna. A számítástechnika újabb fejlődése a szűkebb értelemben vett tesztvizsgálatok (intelligencia- és tantárgytesztek) előtt is olyan új távlatokat nyit, amelyek nélkülözhetetlenné teszik a probabilisztikus tesztmodelleket. A nagyon nagy feladatkészlettel dolgozó *tesztbankoknak,* feladattáraknak a „méretre szabott", *adaptív tesztelés*ben való felhasználása ugyanis a klasszikus tesztelmélet alapján (a populációfüggőség miatt) kielégítően nem oldható meg. A nehézségi szintek populációtól független meghatározhatósága, az egyén képességeinek megfelelő (adaptált) itemkészlet kiválasztása, az itemkészletek egybevethetősége (chaining) viszont olyan követelmények, amelyek éppen a probabilisztikus tesztmodellek alapján teljesülhetnek.

Természetesen a probabilisztikus tesztmodell sem csodaszer. (Éppen FISCHER, 1983, 664. o., figyelmeztet például az adaptív tesztekkel kapcsolatban a probabilisztikus modellek alapján is felmerülő nehézségekre.) A modern tesztelmélet fő törekvése az, hogy a tesztvizsgálatokban – mint mindenütt a tudományos kutatásban – ne vaktában alkalmazzuk a különböző módszereket és eljárásokat, hanem legyünk tisztában feltételeikkel, korlátaikkal. Alkalmazásuk ekkor járulhat hozzá eredményesen ah-

hoz, ami minden tudományos kutatásnak és a gyakorlatnak is közös igénye: a valóság helyes megismeréséhez. Ez a követelmény a hagyományos tesztek és a modern probabilisztikus modellek alkalmazására egyaránt vonatkozik.

Irodalom *

ANASTASI, A., 1976, *Psychological Testing,* Macmillan, New York, London.

EIGLER, G. und STRAKA G.A., 1978, *Mastery Learning. Lernerfolg für jeden?* Urban und Schwarzenberg, München.

EYSENCK, H.J., 1980, *Intelligenz, Struktur und Messen,* Springer, Berlin, Heidelberg, New York.

FISCHER, G.H., 1974, *Einführung in die Theorie psychologischer Tests,* Huber, Bern.

FISCHER, G.H., 1983, Neuere Testtheorie, In: *Messen und Testen.* Bd 3 der Serie Forschungsmethoden der Psychologie der Enzyklopädie der Psychologie, Dr. C.J. Hogrefe, Göttingen, 604–691.

GUTJAHR, W., 1974, *Die Messung psychischer Eigenschaften,* V. d. Wissenschaften, Berlin.

GULLIKSEN, H., 1950, *Theory of mental tests,* Wiley, New York.

HAJTMAN Béla, 1968, *Bevezetés a matematikai statisztikába pszichológusok számára,* Akadémiai Kiadó, Budapest.

KUBINGER, K.D., 1980, Die Bestimmung der Effektivität universitärer Lehre unter Verwendung des Linearen-logistischen Test-Modells von Fischer, *Archiv für Psychologie, 133,* 69–79.

KUN Miklós, SZEGEDI Márton (szerk.), 1971, *Az intelligencia mérése,* Akadémiai Kiadó, Budapest.

LADÁNYI János, CSANÁDY Gábor, 1983, *Szelekció az általános iskolában,* Magvető, Budapest.

LAZARSFELD, P.F., (1966), 1973, Latent Structure Analysis and Test Theory, In: LAZARSFELD, P.F., and HENRY, N.W. (eds.), id. mű.

LAZARSFELD, P.F., and HENRY, N.W. (eds.), 1966, *Readings in Mathematical Social Science,* Chicago, oroszul: Moszkva, 1973.

LAZARSFELD, P.F., and HENRY, N.W., 1968, *Latent Structure Analysis,* Houghton-Mifflin, Boston.

LIENERT, G.A., 1961, *Testaufbau und Testanalyse,* Weinheim.

LORD, F.M., (1966) 1973, The Relation of a Test Score to the Trait Underlying the Test. In: LAZARSFELD, P.F., and HENRY N.W. (eds.), id. mű.

LORD, F.M., NOVICK, M.R., 1968, *Statistical theories of mental test scores,* Addison-Wesley, Reading/Mass.

RASCH, G., 1960, Probabilistic models for some intelligence and attainment tests, Paedagogiske Institut, Kopenhagen.

RASCH, G., (1966) 1973, An individualistic approach to item analysis. In: LAZARSFELD, P.F., and HENRY, N.W. (eds.), id. mű.

*Részletes bibliográfiát lásd FISCHER, G.H., 1974, 1983.

STELZL, I., 1982, *Fehler und Fallen der Statistik,* Huber, Bern.
WAKENHUT, R., 1974, *Messung gesellschaftlichpolitischer Einstellungen,* Huber, Bern.
WILLMOT, A.S., and FOWLES, D.E., 1974, *The objective interpretation of test performance. The Rasch modell applied,* NFER Publ. Co.

GYÖRGY HORVÁTH

TEST-THEORY: PROBLEMS AND PERSPECTIVES

The study departs with a brief survey of the events leading up to the more recent efforts of test-theories: the road from "classical theories" to probability-models. It outlines the development, from the beginnings of the "latent-structure" trend worked out by Paul Lazarsfeld through L. Guttman's "perfect scale", leading to Rasch's model and the work of G. Fischer.

Then it goes on telling the premises of classical test-theory (the axioms formulated by Gulliksen) and tells the conclusions that could be derived from the basic equation of classical test-theory (test score = true score + error) and from its premises. The main concern of classical test-theory is the problem of the accuracy of measurement (reliability) and of the accuracy of prediction (validity). The most important momentum in the modern criticism of classical test-theory is the demonstration of the test characteristics being dependent on the population measured, which this article is intended to show in concern of reliability.

Then it surveys the not-yet-cleared up problem about scale features (the intervallum-scale and the premise of normal distribution). The study shows how these problems can lead to arte-fact, in the case of measuring changes for instance. After summarising the criticisism of classical test-theory it draws attention on some possible errors in practice that stem from ignoring of the premises of right the same classical test-theory. In application, the relativity of test-norms due to the mode of standardization for instance is not yet always being considered along with shifts in interpretation due to differing SD-values. Applicators not always consider uncertainties necessarily stemming from errors in measurement when interpreting differences of testpoint-values.

Among latent structure-models, first Guttman's deterministic model is introduced for it had in many respects prepared ground for probability-models. Concerning Lazarsfeld's probability-model the role of accounting equations is considered and also such important concept like local stochastic independence, sufficient statistics and itemcharacteristics. Among probability-models the principles and the probabilistic attitude of the Rasch-model is dealt with some details then the different ways of the description are touched upon. The verifiability of the model is emphasized, then the principles of each model probes are told. After all these, other probability-models are briefly considered, among them the most commonly used binominal model as well. Closing the study a brief survey of possible applications and examples follow.