Bírálat

Molnár Gyöngyvér „Technológiaalapú tesztelés az oktatásban: a problémamegoldó képesség fejlődésének értékelése”című akadémiai doktori értekezéséről

Molnár Gyöngyvér modern tesztelméleti alapokra épülő, a tanulók értékelésével, ezen belül az adaptív teszteléssel, illetve a problémamegoldó képesség fejlettségének mérésével kapcsolatos értekezése a magyar neveléstudomány számára az egyik legjelentősebb kérdést ragadta meg. Nagy feladata, sőt, adóssága a hazai neveléstudományi kutatói társadalomnak, a fejlesztőknek és a pedagógiai gyakorlatnak is a modern tesztelméleti alapokra épített mérési, értékelési kultúra elsajátítása, elterjesztése. Ezért a témaválasztás kiválónak tekinthető. A felvetett kérdéseknek, vizsgálati problémáknak Molnár Gyöngyvér – nem jelent nagy kockázatot a kijelentés megfogalmazása – első számú hazai kutatója, aki a nem könnyű szakmai kérdések, kutatási eredmények hazai megismertetésében is jelentős szerepet vállal.

Bírálóként az alábbiakban számos – elsősorban értelmezési, elvi jellegű – problémát elemzek, azonban szeretném előre bocsátani végső érékelésem summázatát: Molnár Gyöngyvér értekezését magas színvonalúnak értékelem, nyilvános vitára bocsátását messzemenően támogatom. Néhol meglehetősen élesnek tekinthető, alábbi kritikai megállapításaim ellenére is úgy látom, hogy Molnár Gyöngyvér dolgozata kiváló, a jelölt rendkívül magas színvonalú szakmai, tudományos tevékenységének bizonyítéka.

Az értekezés egésze azt mutatja, hogy Molnár Gyöngyvér a tesztekkel történő mérések, az e mérések eredményeire épülő tudományos kutatások rendkívül alapos ismerője, értője, és gyakorlott kivitelezője. Meggyőződésem, hogy ma Magyarországon ő az, aki ezen a területen a legnagyobb tudományos kutatási gyakorlattal, a legszélesebb ismeretbázissal rendelkezik, és képes kreatív módon új kérdéseket feltéve a terület tudományos továbbfejlesztésére, és erről tanúskodik az értekezés is. Elméletileg jól megalapozott, magas színvonalú, és igencsak nagy volumenű kutatómunka részletei bontakoznak ki.

Az értekezés jelentős tudományos eredményeket tartalmaz, elsősorban a problémamegoldó képesség fejlettségének modern tesztelméleti eszközökkel történő vizsgálatával, illetve az adaptív tesztelés részleteinek elemzésével kapcsolatban. Az empirikus kutatás adatai hitelesek, az eredmények meggyőzőek. A szerző kutatási eredményei közül mindenképpen új tudományos eredményekként fogadhatók el a következők:

- Kisiskolás diákok körében az adaptív tesztelés kutatásokban és értékelési folyamtokban történő alkalmazásához szükséges eszközhasználati képességek megfelelő fejlettségi szintjének vizsgálatával kapcsolatos eredmények.
- Az iskolakezdés szempontjából fontos készségek mérése technológiaalapú megvalósíthatóságának részbeni igazolása, illetve a problémák tisztázása.
- A dinamikus problémamegoldó képesség fejlődésével kapcsolatos összefüggések feltárása technológiaalapú módszertani megoldásokkal.
- A problémamegoldó stratégiák logfile-ok elemzésén alapuló feltárása.

A bírálatban, a továbbiakban azonban inkább a kritika lesz domináns. Ugyanis bizonyos részleteiben még ez a munka is rendelkezik hiányosságokkal, problematikus értelmezésekre épülnek néhány helyen az érvelések. A bírálónak persze kötelessége ezeket szóvá tenni, de azért is teszem, hogy a felvetődő kérdésekben vitákat kezdeményezzek, a célom sokkal inkább egy

tudományos diskurzus kialakítása, és egyáltalán nem az akadémiai doktori cím elnyerésére való alkalmasság megkérdőjelezése.

Problémák a méréselméleti megalapozottsággal

Molnár Gyöngyvér értekezésében nem hivatkozik a neveléstudományi mérések reprezentációs méréselméleti, matematikai megalapozására. Ez még önmagában nem szakmai hiba, a nemzetközi tudományos életben nincs szükség már arra, hogy a kutatások legmélyebb elméleti megalapozásáig visszanyúljunk. A neveléstudományi mérések esetében azonban két okból lenne jelentősége annak, ha a hozzáértők itthon ezt mégis megtennék: (1) Magyarországon ma még csak szűk körben ismert alapozásról, a kutatásokban és publikációkban egyáltalán nem szereplő témáról van szó, (2) a teszteket alkalmazó kutatások esetén – és ez igaz Molnár Gyöngyvér munkájára is – jó néhány elméleti megállapítás nincs összhangban magával a reprezentációs méréselmélettel. Utóbbit nézzük konkrétabban is!

Fontos megállapítás a 16. oldalon, hogy „…a teszt-összpontszámmal való korreláció jól megmutatja, illik-e egy item a tesztbe, ugyanazt méri-e, mint a többi". Itt egy széles körben elterjedt eljárásról van szó. Egyszerűen fogalmazva: ha a kutató azt látja, hogy kicsi a teszt-item korreláció, e tényből azt a következtetést vonja le, hogy az item nem ugyanazt méri, mint a teszt egésze, és így a szóban lévő itemet kihagyja a vizsgálatból. Az alacsony teszt-item korrelációnak azonban más oka is lehet, az tudniillik, hogy a vizsgált képesség fejlettsége nem mérhető, valójában ez a fejlettség nem is létezik. Egy képesség fejlettsége akkor mérhető a klasszikus tesztelmélet szerint, ha a képességhez tartozó feladatokból csupa azonos eredetű (congeneric) teszt állítható elő (Jöreskog 1971; Steyer 2001). E fogalom azt jelenti, hogy bármely két teszt valódi pontszámai (a szerezhető pontszámok várható értékei) a vizsgált populáción egymással pozitív lineáris kapcsolatban vannak. Ez az elvárás nagyon természetes is, hiszen ha a mérés intervallumskálákra alapozott kell legyen, akkor a minimum, hogy ez a linearitás érvényesül. Mindebből levezethető, hogy ha ebben az értelemben mérhető képességfejlettségről van szó, akkor a tényleges mérés során kapott itempontszámok és a mért tesztpontszámok közötti korrelációs együtthatóknak magasaknak kell lenniük. Ez áll annak a hátterében, hogy az alacsony teszt-item korrelációból még nem feltétlenül az következik, hogy az item „rossz". Az is lehet az ok, hogy a képesség fejlettsége nem mérhető, a képességhez tartozó feladatokból előállítható tesztek nem azonos eredetűek. Az itemszelekció abban az értelemben egy „káros művelet", hogy elrejti a nem mérhetőséggel kapcsolatos problémát. Olyan teszt konstrukciójához vezet, amelynek a résztesztjei azonos eredetűek, a Cronbach-alfa értéke magas, ezért klasszikus tesztelméleti értelemben a vizsgálat lefolytatható. Csak az a baj, hogy nem azt mérjük, amit eredetileg szerettünk volna, hanem annak csak egy valamilyen dimenzióját.

De a klasszikus tesztelméleti értelemben vett – legalább az amúgy nem ismert dimenzió mérése szempontjából – korrekt méréssel van egy további baj is. Az, hogy a reprezentációs méréselmélet szempontjából illegitim. A reprezentációs méréselmélet (Luce és Suppes 2002; Narens 1981) megköveteli, hogy a mért dolgok halmazán érvényesüljenek olyan relációk, amelyek lehetővé teszik, hogy e halmaz és a valós számok valamely részhalmaza között létezzen homomorf leképezés (a halmazon érvényes relációknak megfelelő, a valós számokon érvényes relációkkal). Mivel a számításainkban matematikai statisztikai, és a rendstatisztikáknál rendszerint magasabb szintű módszereket akarunk alkalmazni, ezért általában az a minimális elvárás, hogy az adott, a mérendő dolgokat tartalmazó halmazon úgynevezett különbségi struktúra létezzen (Roberts 1979). Egyszerűen – nem részletezve a matematikai összefüggéseket – értelmes kijelentésnek kell annak lenni, hogy (konkrétan fogalmazva) két tanuló képességfejlettsége közötti egyfajta „távolság" nagyobb, mint másik két tanuló képességfejlettsége közötti „távolság". Ha ez a helyzet, akkor a mérés intervallumskálát eredményez. Ha a kutatók nem specifikálnak egy ilyen struktúrát, akkor a mérés illegitim, legalábbis a reprezentációs méréselmélet szempontjából. Nem születik intervallumskála, és tudjuk, hogy ez a tény teljes mértékben bizonytalanná tesz minden kicsit is bonyolult statisztikai vizsgálatot. A klasszikus tesztelmélet azért illegitim, mert soha nem született megoldás a reprezentációs méréselméleti megalapozására.

Két probléma van tehát: a klasszikus tesztelmélet egyrészt nem felel meg a reprezentációs méréselméletnek, másrészt olyan feladatot fogalmaz meg részben – a nagyon komplex, összetett, sok-sok dimenzióból álló képességek fejlettségének mérését –, ami nem teljesíthető, a mérés lehetetlen, legitim skálák nem alakulhatnak ki. E problémák közül a modern tesztelmélet csak az elsőt oldja meg. A modern tesztelmélet (legalábbis annak egyparaméteres Rasch-modellje) megfelel a reprezentációs méréselméletnek. A probléma a második feltétellel ott van, hogy a komplex képességek esetén a modern tesztelmélet sem nyújt megoldást a mérésre, mert a komplex képességek fejlettségeivel kapcsolatban a Rasch-modell nem működik. Erre vissza kell térnem a dolgozat azon részeinek bírálatánál, amely részek a problémamegoldó gondolkodás fejlettségének vizsgálatáról szólnak.

A reprezentációs méréselméletre való hivatkozás hiányára vonatkozóan nagyon fontos a következő megjegyzés az értekezésben:

> Két azonos képességet mérő teszt esetén … a tesztek nyerspont-értéke csak a diákok egymáshoz viszonyított sorrendjéről ad információt, de a közöttük lévő képességszintbeli távolságról nem, miután az annak függvényében változik, hogy könnyű vagy nehéz tesztet oldottak meg a diákok. (18. o.)

Először is vonatkoztassunk el attól a fogalmazási hibától, hogy itt nyilván nem két, hanem csak egy tesztről van szó. A klasszikus tesztelméleti szabályoknak megfelelő eljárásban sajnos még az sem biztos, hogy a nyers pontszámok, vagy a százalékos értékek a felmértek sorrendjéről tájékoztatnak. A komplex képességek esetén ordinális, sőt még nominális skálára sem számíthatunk. Nem léteznek ugyanis a reprezentációs méréselméletben „előírt” struktúrák. Ilyen skálák nem is alkothatók a komplex képességek esetében. Ha van a képességhez tartozó feladatoknak egy részhalmaza, amelyet a vizsgált populáció egy csoportjába tartozók nagy valószínűséggel tudnak megoldani helyesen, míg a feladatok egy másik halmazával sokkal nehezebben birkóznak meg, illetve egy másik csoport a populáción belül ezzel éppen fordítva van, akkor tudunk szerkeszteni olyan tesztet, amelyben a két csoportba tartozók nagyjából azonos eredményt érnek el, de tudunk olyat is, amelyben az egyik csoport tagjai lényegesen jobban teljesítenek, mint a másik csoport tagjai, illetve ezt fordítva is megtehetjük. A sorrend nem ugyanaz a különböző tesztek esetén, de még az azonosság is tesztfüggő. Az viszont nem igaz, hogy a klasszikus tesztelméletben problémát jelent a különböző nehézségű tesztekkel szerezhető pontszámok közötti különbség. A hossz mérésekor is számtalan különböző skála létezik, más a televízió képernyő átlója hosszának a mértéke, ha cm-ben és ha inch-ben adom meg, de ettől a hosszmérést nem tekintjük hibásnak. Az abszolút skála használatának kivételével (pl. darabszám) a legitim mérések esetén végtelen sok skála van, ám ha a mérés megfelel a reprezentációs méréselméletnek, ez nem probléma. Az értékek megfelelő transzformációval átszámíthatók egymásba.

A szöveg így folytatódik: „Egy könnyebb és egy nehezebb, ugyanazon képességet mérő teszten elért összpontszámok közötti kapcsolat nem lineáris, két különböző nehézségű teszten nyújtott nyerspont-alapú teljesítmény direkt összehasonlítása nem releváns” (18. o.). Ha a mérés megfelelne a klasszikus tesztelmélet szigorú kívánalmainak, vagyis az adott képesség fejlettségét mérő tesztek azonos eredetűek lennének, akkor igenis lineáris lenne a kapcsolat. Ezt nem befolyásolja a tesztek könnyebb és nehezebb jellege. Mint már jeleztem, a komplex képességek esetén ez lehetetlen. A nyerspont alapú pontszámok (precízebben: a háttérben álló, látens értékek) igenis összehasonlíthatók lennének, ha a mérés valóban megfelelne az elvárásoknak, a két teszt azonos eredetű lenne, és ismernénk az egyik teszt pontszámainak a másikba való átszámítására szolgáló lineáris összefüggést. Ez ugyanaz, mint, hogy ismerjük a Celsius hőmérsékleti skála Fahrenheit skálára történő átszámításának képletét, a két skála különböző értékei a hőmérséklet mérése során semmilyen problémát nem jelentenek.

A klasszikus tesztelmélet értékelésével kapcsolatos problémák

Valójában már az előző pontban is jeleztem néhány olyan értelmezési problémát, amely a klasszikus tesztelmélet szerepével, koncepcionális alapjaival, ezek értekezésben való megjelenésével kapcsolatosak. Vannak azonban továbbiak is. A következő mondat tartalma sem felel meg a „tiszta" értelmezéseknek:

> A klasszikus tesztelmélet eszközrendszerével a diákok képességszintjének meghatározása jelentős mértékben függ a kutatás során alkalmazott teszt(ek) nehézségi szintjétől (18. o.).

A konkrét teszthez kötött pontszámok természetesen valóban függnek a konkrét teszt nehézségi szintjétől, mint ahogyan a TV képernyő átlója hosszának a mérőszáma is függ attól, hogy milyen mércével, cm vagy inch beosztásúval mérem. A mondat elsődleges tartalma tehát megfelel a valóságnak, csakhogy nyilvánvalóan azért íródott le, mert a szerző a benne lévő tartalmat problematikusnak tartja. Valójában azonban nem az, ha a klasszikus tesztelmélet szigorú követelményeinek teljesülését várjuk el.

És még mindig ugyanez a probléma a következő szövegrészben is:

> A nyers- vagy százalékpontok használatának további problémája a teszt itemei nehézségi szintjeinek és a diákok képességszintjeinek összekapcsolása. Egy ideális mérés során elvárjuk, hogy ha egy diák pl. 55 pontot ér el 100 pontból, akkor meg tudjuk mondani, hogy mit tud, az adott képesség fejlődésének milyen stádiumában van, mi várható el tőle. Ha nyers adatokat használunk a tanulók képességszintjének és az itemek nehézségi szintjének meghatározásakor, nem egyértelmű, hogy hogyan kapcsoljuk össze a két skálát. (18. o.).

Először is, nem triviális, hogy a képességfejlettségeket és a feladatnehézségeket össze kell kötni. A modern tesztelmélet a problémát valóban úgy oldja meg, hogy e két mennyiség egy skálára kerül, de ez nem szükségszerű. A klasszikus tesztelmélet – legalábbis a megalapozásában – nem is szól feladatnehézségekről. A képességfejlettség mértékét az adott teszt esetén szerezhető pontszámok várható értékeként értelmezi, és akkor tekinti mérhetőnek, ha a képességhez tartozó összes lehetséges teszt azonos eredetű (ahogy fentebb már szerepelt). A klasszikus tesztelmélet korrekt eljárást javasol, amennyiben igencsak korrekten definiálja a képességfejlettséget, e definícióban nem használja fel a feladatnehézség semmilyen értelmezését, és megadja azt is, hogy mikor tekinti a képességfejlettséget mérhetőnek. Ha a klasszikus tesztelméletnek lenne reprezentációs méréselméleti megalapozottsága, vagyis a képességfejlettségek halmazán adott lenne egy különbségi struktúra, és a tesztpontszámok várható értékét rendelve a képességfejlettségekhez ez egy homomorf leképezés lenne, akkor az így felépített elmélet következetes módon kielégítene minden méréselméleti igényt. A klasszikus tesztelmélettel nem az a baj, hogy a képességfejlettségeket és a feladatnehézségeket nem egy skálán helyezi el, hanem az, hogy valójában skála sincs, hiszen nem létezik az elmélet reprezentációs méréselméleti megalapozása.

A klasszikus tesztelmélet bírálatát tekintve a szerző újra és újra visszatér ugyanahhoz az állításhoz, de ez az állítás nem megalapozott. A 19. oldalon is:

> A klasszikus tesztelmélet eszközrendszerével kizárólag azon feladatok és tesztek eredményeiről, megoldottságáról beszélhetünk, és csak azokat elemezhetjük, azon teszteredményekből vonhatunk le következtetéseket, amelyeket a valóságban is megoldott a diák. Arról nem mondhatunk semmit, hogy ugyanazon diák, esetleg ugyanazon a vizsgált képességterületen egy könnyebb, vagy egy nehezebb teszten hogyan teljesített volna.

Ha egy mérés valóban megfelel a klasszikus tesztelmélet követelményeinek, akkor az adott képességhez tartozó feladatokból csupa azonos eredetű teszt állítható elő. A tesztekhez tartozó látens, elméleti képességfejlettségek, vagyis a pontszámok várható értékei tesztpáronként pozitív

lineáris kapcsolatban állnak egymással, és megfelelő módszerekkel a lineáris kapcsolat paraméterei jól becsülhetők. Ebből az következik, hogy elvileg bármely teszt pontszámait átszámíthatjuk bármely más teszt pontszámaivá. Nem itt van a baj. A baj ott van, hogy a komplex képességek esetén a klasszikus tesztelmélet követelményeinek való megfelelés nem teljesül, illetve ezen a módon nem állnak elő intervallumskálák.

Problémák a modern tesztelmélet bemutatásával

A 17. oldalon találkozunk először a modern tesztelméletnek (vagy angol nyelvterületen leginkább elterjedt névvel IRT, Item Response Theory) valószínűségi tesztelméletként való megnevezésével. Ezzel a névvel a nemzetközi szakirodalomban ritkán találkozunk, nálunk azonban elterjedt. Bár megnevezéseken nem nagyon érdemes vitatkozni, mégis szóvá teszem, mert ez egy érdemi kérdés is, hogy a modern és a klasszikus tesztelmélet között a valószínűségi jelleg tekintetében egyáltalán nincs különbség. A klasszikus tesztelmélet ugyanúgy valószínűségi alapfogalmakra épül, ahogyan az IRT modellek elméletei. A megnevezést nem tartom jónak.

Ennek megfelelően nehezen értelmezhető a 18. oldalon szereplő leírás:

> A valószínűségi tesztelmélet a mérés során elkövetett hibát és az itemek tulajdonságait más módon, nem determinisztikusan, hanem valószínűségi alapon kezeli.

A klasszikus tesztelmélet is valószínűségi alapon kezeli a mérési hibát, hiszen magát a mérési folyamatot sztochasztikusként értelmezi, a felmerülő mennyiségek várható értékeiről, varianciáiról, korrelációiról beszél. A modern tesztelméletben az itemek paraméterei, de a felmért személyek képességparaméterei is abban az értelemben determinisztikusak, hogy a mérési modellben eleve adottaknak, állandóknak, egyfajta látens értékeknek tekintjük őket. Ami valószínűségi, az a képesség megnyilvánulása, az tudniillik, hogy a vizsgált személyek valamekkora, nem 0 és nem 1 valószínűséggel tudják jól megoldani az egyes feladatokat.

Problematikus megállapítások szerepelnek a következő szövegrészletben is:

> A teszt, illetve itemek nehézségi szintjétől független képességszint-meghatározás előfeltétele egy olyan nyerspont-transzformáció, egy olyan matematikai függvény alkalmazása, ami megszűnteti a teszt nehézségétől függő képességeloszlást. Erre alkalmas matematikai összefüggést biztosít a valószínűségi tesztmodellek közé sorolható Rasch modellben használt logisztikus függvény. A Rasch modell logisztikus transzformációja a nyers adatokat egy olyan skálára transzformálja, ami nemcsak a diákok közötti sorrendet, hanem a diákok közötti távolságok nagyságát is megőrzi. (18. o.).

A Rasch-modell alkalmazása során a képességfejlettség becslések meghatározására valóban létezik olyan módszer, amely a nyerspontszámok transzformációját végzi (PROX eljárás). Ez azonban egy a létező módszerek közül, a legkevésbé pontos (bár meglepően jól használható adatokkal szolgál), mert egyébként összetett statisztikai eljárások állnak rendelkezésre a pontosabb becslésekhez (legkisebb négyzetek módszere, maximum likelihood becslés, stb.). A lényeg magyarázata egy perifériális tényezővel nem sikerülhet, valójában a Rasch-modell lényegét nem a transzformációval érdemes megragadni. (Csak fogalmazási probléma, hogy természetesen nem a képességeloszlást kell megszűntetni, hanem a képességeloszlás tesztnehézségtől való függését.) Molnár Gyöngyvér ezen a ponton sajnos nem ír explicit definíciókat. A leírás azt sejteti, hogy előbb egy matematikai transzformációval a nyerspontszámokból előállítjuk a képességfejlettségeket, majd egy adott feladat esetén megnézzük, hogy milyen képességfejlettséggel rendelkező személy tudja azt 0,5 valószínűséggel megoldani helyesen, és ennek a személynek a képességfejlettsége lesz a feladat nehézsége. Ez így nem definiálás, és nem azért, mert nem ismerjük meg a matematikai transzformáció képletét, hanem azért, mert itt Molnár Gyöngyvér a mért értékekből a becslések előállításának egyfajta algoritmusát adja meg, és nem a látens képességfejlettségek és

feladatnehézségek eredeti definícióját. Hogy ne legyek vádolható azzal, hogy a levegőbe beszélek, ide írom a definíciót:

> Legyen adott embereknek egy populációja, és egy adott képességhez tartozó feladatok halmaza. A képesség az adott populáción megfelel a Rasch-modellnek, amennyiben a személyekhez hozzárendelhetők úgy képességfejlettségeknek nevezett valós számok, és a feladatokhoz hozzárendelhetők úgy feladatnehézségeknek nevezett valós számok, hogy bármely feladat adott személy általi helyes megoldásának valószínűsége, az adott feladat nehézségparamétere, és a kiválasztott személy képességfejlettsége között a következő összefüggés érvényes:
>
> $$p_{ij} = \frac{e^{\beta_i - \delta_j}}{1 + e^{\beta_i - \delta_j}},$$
>
> ahol p_{ij} a j-edik feladat i-edik személy általi jó megoldásának valószínűsége, β_i az i-edik személy képességfejlettsége, és δ_j a j-edik feladat nehézsége.

Ebben a definícióban szó sincs mért értékekről, hiszen az itt szereplő mennyiségeket először elméletben kell definiálni, és csak utána merülhet fel a kérdés, hogy az így definiált mennyiségek miképpen mérhetők. Közvetlenül egyébként semelyik nem mérhető, ám a ténylegesen meghatározható adatokból (sikerült, vagy nem sikerült megoldani a teszt feladatait), becsülhetjük a szóban forgó értékeket (akár a PROX eljárással is, de ez már részletkérdés).

Sajnos a következő szövegrészletben szereplő állítás sem felel meg a matematikai háttérnek:

> … tesztfüggő, hogy az adott elemzés során milyen távol van egymástól két ember képességparamétere a képességskálán. Egy magasabb diszkrimináló erővel rendelkező teszt jobban széthúzza, jobban diszkriminálja a személyeket, mint egy, az adott mintát kevésbé diszkrimináló feladatlap. (20. o.)

Az egyparaméteres Rasch-modellnek megfelelő skálák abban különböznek egymástól, hogy az egyiken mérhető képességfejlettségekhez és feladatnehézségekhez egy állandót kell csak hozzáadni, hogy megkapjuk a másik skálán mérhető képességfejlettségeket illetve feladatnehézségeket. A Rasch-modell képlete ennyi szabadságot ad. De ez azt jelenti, hogy két ember képességfejlettsége közötti különbség bármely skálán ugyanannyi. Legyen az egyik skálán két személy képességfejlettsége β_1 és β_2, egy másik skálán (másik teszttel mérve) β_1' és β_2'. A két skála úgy függ össze, hogy van egy d valós szám, amellyel $\beta_1' = \beta_1 + d$, és $\beta_2' = \beta_2 + d$. Így $\beta_1 - \beta_2 = \beta_1' - \beta_2'$. Ha két teszt közül az egyik jobban differenciál, mint a másik, akkor kettőjük közül legalább az egyik már nem felel meg az egyparaméteres Rasch-modellnek, és már nem felel meg annak a Georg Raschtól származó követelménynek, amelyet a szerző is egyébként (nagy örömömre) a 18. oldalon idézett.

A 31. oldalon a szerző arról ír, hogy az adaptív tesztelés számára miképpen készítünk elő tesztfeladatokat. Kiemeli a Rasch-modell jelentőségét, amennyiben az lehetővé teszi, hogy „bemérjük" a feladatok paramétereit. A paraméterek között azonban szerepel a diszkriminációs index is. A feladatok diszkriminációs indexe a két- és háromparaméteres Rasch-modell alkalmazása során válik fontossá, azonban ezek a mérések már nem felelnek meg a szigorúbb elvárásoknak. A 18. oldalon a szerző (ezt már méltattam) idéz Georg Rasch-tól, aki a modelljét alkalmazó mérésekkel szembeni elvárást igen pontosan leírta. Ha egyszerre alkalmazunk (akár PP tesztben, akár adaptív tesztelés során) különböző diszkriminációs paraméterrel rendelkező feladatokat, akkor az item jelleggörbék átmetszővé válnak, és innentől elveszítettük a szilárd talajt a lábunk alól. Már nem lesz igaz az, hogy ha egy feladatot valaki nagyobb valószínűséggel tud helyesen megoldani, mint egy másik feladatot, akkor ez minden a vizsgált populációba tartozó személlyel így lesz. A metszéspontnál jellemző képességfejlettségnél kisebb fejlettségű személyek az egyik feladatot nagyobb valószínűséggel oldják meg helyesen, mint a másikat, a többiek viszont fordítva lesznek ezzel. A tesztekkel való mérés két alapelve közül az egyik, a monotonitás alapvetően sérül.

Az adaptív tesztelésről

Jó lett volna bemutatni, hogy a PP tesztek alkalmazása szerint milyen hibákkal kell számolnunk. És itt nem a mintán becsülhető hibáról van szó, hiszen annak nagyságáról a reliabilitás becslése elég jól tájékoztat, hanem a felmért egyénekre jellemző hibákról. A klasszikus tesztelmélet kiindulópontjaként tisztelt $X = T + E$ összefüggés bármilyen mérésre, ezért a modern tesztelméleti alapon zajlókra is érvényes, ezért a hiba nagysága itt is érdekes. De ez a kérdés először egyetlen felmért személy esetén merül fel. Mekkora a hiba egyetlen tesztfelvételnél? Különösen a „nagy mérések" (PISA, TIMSS, OKM, stb.) esetén vannak is arra nézve adatok, vagy számíthatók ilyenek, hogy milyen lehet személyenként a hiba. Pontosabban, pl. az OKM-ben az egyéni mérések standard hibájára kapunk egy becsült értéket. Érdemes lett volna ilyen adatokat bemutatni. Az OKM esetén az adatok szemrevételezése alapján azt állapíthatjuk meg, hogy az egyéni hibák igencsak jelentősek. Egy tipikus példa: ha egy tanuló 1600 pontot ért el a teszten, akkor valódi pontszáma 2/3 valószínűséggel esik az (1540; 1660) intervallumba, és 1/3 valószínűséggel van azon kívül (a becsült standard hibák átlaga nagyon stabilan 60 ponthoz közeli érték). Ez az oka annak, hogy sem az OKM tesztjeinek eredményeit (de véleményem szerint semmilyen PP teszt eredményeit) egyéni értékelésre nem szabad használni. Az adaptív mérés óriási jelentősége éppen abban állna, ha beváltaná a hozzá fűzött reményeket, hogy ezt a jelentős hibát számottevően csökkenthetjük, amivel akár egyéni értékelésre is alkalmassá tehetjük a teszteket. Ezt az összefüggést jó lett volna alaposabban, akár adatokkal is szemléltetve bemutatni.

A mérési pontossággal kapcsolatos korábbi megjegyzésem fényében kritikával kell illetnem a következő gondolatot is:

> Ha a tanuló részt vett már korábbi tesztelésben, ahol a teszt az adott feladatbank feladataiból került összeállításra, akkor a korábbi teljesítménye összevethető aktuális eredményével, még akkor is, ha összességében minden egyes alkalommal más itemeket oldott meg. (31. o.)

A teljesítményre ez az állítás nem állja meg a helyét. Amiket össze lehet hasonlítani, az a valóságos, a háttérben álló, vagy a szakirodalomban leggyakrabban használt szóval látensnek nevezett két képességfejlettség. A klasszikus terminológia, jelölés szerint a T értékek hasonlíthatók össze, és nem az X-ek, vagyis a tényleges teszteredmények. Egy példával illusztrálom. Tegyük fel, hogy van egy tanuló, akinek a látens képességfejlettsége, amit a szövegértés teszt mér (becsül) az OKM-ben hatodik osztályban 1500 pont volt, nyolcadik osztályban pedig 1600 pont. Tipikus tanuló, hiszen 100 pontot fejlődött (és e körüli érték a pontszámnövekedések átlaga), azonban a konkrét teszteredményei természetesen nem 1500 és 1600 pont igen nagy valószínűséggel. Tegyük fel, hogy e mérésekben ez a tanuló először, hatodikban 1555 pontot teljesített, a második, a nyolcadikos mérésben 1545 pontot. Ha csak ezekre a közvetlenül mért értékekre nézünk, azt mondhatnánk (hibásan), hogy a tanuló tudása a vizsgált területen 10 pontot romlott, a tanuló visszafejlődött. Ennek a kijelentésnek semmi köze nincs a valóságos fejlődéséhez. Ahogy már jeleztem korábban, az 1555 és az 1545 pontos teljesítések egyáltalán nem valószínűtlenek. Személyenként az összevetés a teljesítmények alapján nem lehetséges. A modern tesztelméleten alapuló eljárások azt biztosítják, hogy a két különböző mérés esetén a háttérben lévő, közvetlenül nem mérhető, a definícióban meghatározott képességfejlettségek hasonlíthatók össze, mert azonos skálán helyezkednek el. A skálán nem a mért értékek vannak, hanem az elméleti, a látens értékek.

A sikeres iskolakezdés feltételei

A 74. oldalon kezdődik az értekezésnek az a része, amely a sikeres iskolakezdés feltételei rendelkezésre állásának mérésével foglalkozik. E kérdésben felmerül egy fontos elvi kérdés. Mik vajon a sikeres iskolakezdés feltételei, az egyes feltételek tekintetében milyen fejlettségi szint tekinthető valamifajta minimumnak? Tény, ezt a szerző, vagy például Nagy József kutatásai is többször alátámasztották, hogy számos képességet, tágabban tudáselemet tekintve a gyerekek nagyon különböző fejlettséggel érkeznek az iskolába. Az a kérdés, hogy ebből milyen

következtetéseket vonunk le. E következtetések nagyjából kétféle, egymástól alapvetően eltérő elvi alapon nyugszanak. Az egyik paradigma a problémát a deficit fogalmával írja le, a hátrányos szociális helyzetű gyerekek személyisége, tudása deficiteket mutat, ebből következően a különbségeket felzárkóztatással kell csökkenteni, vagyis az elmaradásokkal jellemezhető területeken speciális fejlesztéseket kell végezni. A másik paradigma eleve mást gondol a különbségek természetéről, és ebből kiindulva részben más utat kínál a probléma megoldására. A gyerekek közötti különbségeket nem mennyiségi természetűeknek tartja, hanem minőségieknek. Természetesen egy kiszemelt tudáselem tekintetében mennyiségi különbségről van szó, azonban a szociálisan hátrányos helyzetű gyerekek nem általában maradnak el társaikhoz képest, hanem arról van szó, hogy más tudásszerkezettel bírnak. Vannak a többiekéihez képest fejletlenebb, de vannak fejlettebb képességeik is. A világ bizonyos „szeleteiről" valóban kevesebbet tudnak, de vannak a világnak olyan „szeletei", amelyekben viszont ők rendelkeznek alaposabb tudással. Valóban rosszabbul kommunikálnak verbálisan, de kommunikációjuk összességében (számításba véve más kommunikációs csatornákat és formákat) nem alacsonyabb szintű, mint társaiké. A probléma azért látszik súlyosnak, egyáltalán, problémának, mert az iskola valamifajta ismeretek meglétét várja el, bizonyos képességek magasabb fejlettségét, egyértelműen a verbális kommunikációt helyezi előtérbe, és preferál bizonyos magatartásformákat, értékeket és normákat. Az iskola egyoldalú, előnyben részesíti a fehér középosztály tagjai által birtokolt ismereteket, a gyermekeikben fejlett képességeket, stb. Ezért a fejlesztésnek nem az lenne a feladata, hogy az iskola az általa előtérbe helyezett összetevőket tekintve felzárkóztasson, hanem az, hogy a gyerekek már adott feltételeihez igazodó módon fejlesszen, differenciálással nem csak a bizonyos gyermekcsoportokban már meglévő előfeltételeket vegye figyelembe, hanem mindenfajta a gyermekekben megjelenő értéket. Vagyis az iskolakezdés feltételeinek vizsgálata akkor lenne korrekt, akkor felelne meg az interkulturális nevelés elveinek, ha ezek a bizonyos feltételek egy jóval tágabb halmazt jelentenének, és nem pusztán a fehér középosztályi kultúrához igazodnának.

Az iskolakezdés feltételeinek tesztek alkalmazásával történő felmérése általában a deficit modellt követi. A DIFER mérésben is csupa olyan képességet találunk, amelyekben biztos, hogy a fehér középosztályhoz tartozó gyerekek a fejlettebbek, ezért aztán, amikor kutatási eredményként sikerül kimutatni, hogy a fehér középosztály gyermekei jobban teljesítenek, nagyon nem szabadna csodálkoznunk. A fehér középosztályhoz tartozó gyermekekben nem általában az iskolakezdés feltételei vannak magasabb fejlettségi szinten, hanem azok a képességek, amelyeket az iskola egyoldalúan az iskolakezdés feltételeinek tekint. Természetesen az ezekben mutatkozó különbségek, jellemző összefüggések kimutatása sem haszontalan, de egyelőre igen távol van az „iskolakezdés feltételei probléma" széles szakmai bázist használó megoldásától.

A problémamegoldó képesség fejlettségének mérése

Az értekezés meghatározó jelentőségű fejezetei a problémamegoldó képesség méréséről szóló, a mű kb. felét kitevő részben találhatók. Elismerve a világban a problémamegoldás képessége fejlettségének mérésére született számtalan kezdeményezés, konkrét kutatás, elemzés értékeit, érdemes mégis a témával összefüggésben alapproblémákat felvetni. Meggyőződésem szerint néhány alapkérdés nincs világosan megválaszolva, és ez árnyékot vet a problémamegoldó képesség fejlettségének mindenféle vizsgálatára. A legfőbb alapprobléma egyszerűen megfogalmazható: van-e olyan, hogy problémamegoldó képesség fejlettség, és ami szinte ugyanaz a kérdés: mérhető-e a problémamegoldó képesség fejlettsége. Saját válaszom az, hogy a problémamegoldó képességnek nincs fejlettsége, és ezzel összhangban természetesen akkor nem is mérhető.

Világos, hogy olyan állítást fogalmaztam meg, amely a területen dolgozó minden szakember álláspontjával ellentétes, megkérdőjelezi a PISA felmérés részeként működő problémamegoldó képesség fejlettség mérést, vagyis igencsak alaposan meg kell magyaráznom, miért vélekedek így.

Érvelésem lényege valójában már bírálatom korábbi részeiben is szerepelt. Mérhető értékként akkor beszélhetünk a problémamegoldó képesség fejlettségéről, és annak méréséről, ha az, amit definícióként kimondunk, megfelel a reprezentációs méréselmélet elvárásainak, méghozzá úgy

(figyelve, hogy a szakemberek milyen statisztikai számításokat végeznek a kapott eredményekkel), hogy a fejlettséget jelző látens értékek intervallumskálán helyezkedjenek el. A reprezentációs méréselmélet szerint annak a feltétele, hogy ez sikerüljön, a vizsgált „valamiknek" (jelen esetben a képességfejlettségeknek) különbségi struktúrával kell rendelkezniük. Vezessünk be egy jelölést: legyenek *x* és *y* személyek, és válasszunk ki a képességhez tartozó feladatok közül egyet tetszőlegesen. Jelöljük p_{xy}-nal annak valószínűségét, hogy *x* helyesen, míg *y* rosszul oldja meg a kiválasztott feladatot. A tesztek esetében akkor beszélhetünk arról, hogy a látens képességfejlettségek különbségi struktúrát alkotnak, ha a populációban bármely *a, b, c, d* személyre igaz, hogy ha egy feladat tekintetében $p_{ab}/p_{ba} > p_{cd}/p_{dc}$, akkor ez minden feladattal így van. Bebizonyítható, hogy ez az értelmezés valóban egy ún. különbségi struktúrát hoz létre a képességfejlettségek halmazán, a különbségi struktúrának természetesen van egy axiómarendszere (lásd például Roberts 1979). Az egyparaméteres Rasch-modell felépítése során éppen azt biztosítják a meghatározások, hogy a leírt feltételben szereplő valószínűségek és a képességfejlettségek (és külön a feladatnehézségek) között határozott összefüggés van, amely biztosítja számunkra, hogy az egyparaméteres Rasch-modellel specifikált mérés valóban intervallumskálákhoz vezet. Ez jelenti a modern tesztelmélet (pontosabban az egyparaméteres Rasch-modell) fölényét a klasszikus tesztelmélet felett. Minden képesség esetén az a kérdés, hogy vajon adott populációt tekintve érvényes-e (legalább közelítőleg) a Rasch-modell. A valószínűségekre vonatkozó fenti összefüggésből, de közvetlenül a Rasch-modellben a valószínűségek, a képességfejlettségek és a feladatnehézségek közötti összefüggésből is levezethető a modell érvényesülésének az a szükséges, de nem elégséges feltétele, hogy ha valaki egy feladatot nagyobb valószínűséggel tud megoldani, mint egy másik feladatot, akkor ez a populációban mindenkire igaz kell, hogy legyen, és ez független kell, hogy legyen a feladatpár kiválasztásától. Molnár Gyöngyvér is idézi az ennek megfelelő, még Georg Rasch-tól származó állítást, ezt már korábban is méltattam.

A komplex, nagyon összetett képességekkel az a baj, hogy triviálisan nem teljesíthetik ezt a szükséges feltételt. Az érvet már korábban is leírtam, nem ismétlem meg. Az összetettség, a mi esetünkben a jellegében, tudásigényében a végletekig különböző problémák léte azt eredményezi, hogy a monotonitás a problémamegoldás esetében nem érvényesülhet.

Kérdés, hogy ezt „miért nem tudják" a problémamegoldás képessége fejlettségének mérésével foglalkozó szakemberek, például a PISA szakemberei. Először is, szögezzük le, hogy mindazok, így Molnár Gyöngyvér is, a PISA szakemberei is, akik szakmai alapossággal dolgoznak ezen a területen, nem mérnek hibásan, a mérésekre vonatkozó, általánosan elfogadott követelményeket betartják. Elvégzik a modellek illeszthetőségével kapcsolatos vizsgálatokat, csak olyan teszteket alkalmaznak, amelyek megfelelnek a szigorú követelményeknek. Másról van itt szó, egyáltalán nem hanyag munkáról, vagy szakmai tévedésről. Arról van szó, hogy a tesztek fejlesztése során, elsősorban az itemszelekcióval, valamint a kutatások esetén gyakori megoldással, a „nem jól viselkedő" felmért személyek kizárásával problematikussá válik a mérések validitása. Az az állításom, hogy a problémamegoldás fejlettségét mérő tesztek nem a problémamegoldás fejlettségét mérik, ahogyan a szövegértés tesztek nem a szövegértését, a matematika tesztek nem általában a matematika tudás fejlettségét. Valamit mérnek, ami hozzá is tartozik az általános pszichikus konstrukcióhoz, de nem „fedi le" azt, nem azonos vele. A problémamegoldás fejlettségének mérését szolgáló tesztek a problémamegoldás képessége valamilyen – egyébként közelebbről nem meghatározott – dimenziójának fejlettségét mérik. Ez kevésbé fontos probléma a PISA esetében. Ugyanis a PISA mérésben valójában nem kutatási céllal vizsgáljuk a problémamegoldás képességét. A PISA-ban az érdekel mindenkit, hogy a 15 éves tanulók mennyire felkészültek arra, hogy megoldják a jövőben jelentkező fontos feladatokat. A problémamegoldás mérése során létrejön a teszt (a feladatbank pontosabban), amelynek a feladatai a tesztelés minden szigorú követelményének megfelelnek, egy dimenzióban helyezkednek el. Ám ebből az következik, hogy nem általában mérik a problémamegoldó képesség fejlettségét, hiszen az egyébként sem létezik, ha elfogadható az, amit itt állítok.

Mindez mit mond Molnár Gyöngyvér munkájáról? Molnár Gyöngyvér kutatja a problémamegoldó képességet, kutatásának tárgyát ebben az általánosságban határozza meg. Tesztjei azonban – éppen a módszertani következetesség okán – ténylegesen nem mérik, nem mérhetik az általában vett problémamegoldó képesség fejlettséget. Amit mérnek, az a képesség valamilyen dimenziója. Ha a vállalkozást erre szűkítjük, akkor a munka tökéletes. Hatalmas mennyiségű és biztos kézzel alkalmazott tudás nyilvánul meg benne, olyan mennyiségű munka, amire lehet, hogy nincs is más példa Magyarországon, és mindez lenyűgöző színvonalon. Ha a szerző azt állítja, hogy tesztjei a problémamegoldó képesség fejlettségét mérik, akkor így a munkájával kapcsolatban komoly validitási problémák merülnek fel. Ám ha azt mondja, hogy valójában a problémamegoldó képesség valamely fontos dimenziójának fejlettségét méri, akkor máris korrektté válik az egész tevékenység.

Jelentős kérdések merülnek fel persze. Tényleg fontos az a dimenzió, amit Molnár Gyöngyvér mér? (Például az oktatás szempontjából.) Tekinthetjük valamifajta proxynak? Érdemes kutatni ilyen dimenziókban a problémamegoldó képességet (és nem általában az egész képességet)? Hogyan lehetne más dimenziókat megtalálni? Ha képesek lennénk többféle dimenziót megvizsgálni, annak mi lenne a haszna a gyakorlatban? Ezek persze mind gyakorlati, a témában elérhető eredmények adaptivitásával összefüggésben felmerülő kérdések. Tudományosan valójában „egyszerű” a helyzet: ha lemondunk arról, hogy a problémamegoldó képességet teljes globalitásában akarjuk egyetlen számmal jellemezni, akkor a különálló dimenziók vizsgálata értelmes tudományos feladat, a jelenleg adott elméleti háttérrel és „technológiával” kivitelezhető. Még akkor is, ha a problémamegoldás képességét alkotó dimenziók halmaza beláthatatlan.

Még egy fontos kérdést kell felvetnem. A képességek tudományos vizsgálata során fontos kérdés, hogy magát a képességet miképpen értelmezzük. Az operacionalizálás szempontjából jó gyakorlati megoldás, ha azt mondjuk, hogy bármely képesség esetén elég jól tudunk feladatokat hozzárendelni a konstruktumhoz, feladatok vizsgálata során igen jó biztonsággal meg tudjuk mondani, hogy az igényli-e a képesség közreműködését. A képességek társadalmi konstrukciók, évezredek alatt formálódtak, és csak napjainkat jellemzi az a törekvés, hogy e konstrukciókhoz pszichológiai, sőt, agyi korrelátumokat rendeljünk hozzá. Az esetek nagy részében ez valószínűleg hiábavaló törekvés, a társadalmi konstrukcióként történő értelmezés következetesnek és bármely képesség esetén járható útnak tűnik. Vagyis a képességek feladatokkal, azok halmazaival történő meghatározása kellően jó megoldás. A problémamegoldó képesség definícióiban egyöntetűen szerepel egy elem: az emberek számára azok a feladatok a problémák, amelyekhez nincs azonnali, előhívható algoritmusuk a megoldásra, a megoldás során tudáskonstrukcióra van szükség. Ezt akár megnyugtatónak is tekinthetnénk, hiszen sikerült elég jól elkülöníteni az emberek által megoldott feladatok között a problémáknak nevezetteket. Így azonban egy olyan konstruktumot hoztunk létre, amely személyfüggő. Egy feladat – a feladatok döntő többsége ilyen – lehet az egyik ember számára probléma, míg a másik számára nem az. A problémamegoldó képességnek ez a jellegzetessége jelentős hatással kellene, hogy legyen a vizsgálatokra, a mérésekre, a tesztfejlesztésre. Úgy látom, hogy a problémamegoldással foglalkozó szakemberek tudatosan vagy ösztönösen figyelembe is veszik ezt a helyzetet, és igyekeznek olyan problémákat a tesztekbe helyezni, amelyek szinte biztos, hogy minden felmért személy számára problémák. Ebben azonban soha nem lehetnek biztosak.

Néhány további kritikai észrevétel

Az értekezésben a számítógépek oktatásban játszott általános szerepének elemzése terjedelmét tekintve eltúlzott. Van kapcsolata a disszertáció témájával, de nem tartozik bele.

Zavarba ejtő a következő fogalmazás: „…a PISA-adatok 2000 és 2003 között nemzetközi szinten exponenciális növekedést detektáltak az IKT oktatásban történő jelenlétét illetően” (11. o.). Két pontból hogy lehet megállapítani, hogy a fejlődés exponenciális? Amúgy természetesen jelentős a fejlődés, ez igaz.

Végül: az egyébként jól szerkesztett, jól olvasható írásműben néhány, talán nem elhanyagolható mennységben előforduló nyelvi hiba zavarta a megértést. Néhány példát említek csak:

- A 8. oldalon: „Túlsúlyba kerültek a befektetési-, ingatlan-, biztosítási-, üzleti jellegű szolgáltatást biztosító munkahelyek száma…”.
- Egy másik helyen: „…nem a technológia a cél, hanem az oktatásban jelentkező problémák hatékony megoldására alkalmazzuk azokat…” (13. o.).
- Egy újabb nyelvi hiba: „a technológiaalapú tesztelésre történő átállást segítő kutatások mintája is rendszerint idősebb korosztályokra fókuszálnak” (61. o.).

Összefoglalva: A Molnár Gyöngyvér által bemutatott tudományos kérdések, illetve a rájuk adott válaszok kapcsán néhány fontos szemléleti, elméleti problémát felvetettem az értekezéssel kapcsolatban, és ezeknek a hazai neveléstudományban történő továbbgondolását a tesztekkel végzett kutatómunka továbbfejlődése szempontjából nagyon lényegesnek is tartom, azonban a munka értékeit olyan jelentősnek értékelem, hogy a problémák ellenére is támogatom, hogy az értekezés nyilvános vita tárgya legyen. Meggyőződésem, hogy az általam felvetett problémák jövőbeli megoldásához, az értelmezéseknek a kutatómunkát nagymértékben segítő tisztításához maga Molnár Gyöngyvér is tevékenyen hozzájárul majd. Felkészültsége és tapasztalatai erre maximálisan alkalmassá teszik.

Csömör, 2017. április 2.

Nahalka István

a neveléstudományok kandidátusa

A bírálatban szereplő művek

Jöreskog, K.G. 1971. Statistical Analísis of Sets of Congeneric Tests. *Psychometrica*, *36*(2), 109-133. Az Interneten 2017 január 10-én: https://www.researchgate.net/profile/Karl_Joereskog/publication/24061975_Statistical_Analysis_of_Sets_of_Congeneric_Tests/links/0046352b8b07933bda000000/Statistical-Analysis-of-Sets-of-Congeneric-Tests.pdf

Luce, R.D. és Suppes, P. 2002. Representational Measurement Theory. In Pashler, H. és Wixted, J. (Szerk.) *Stevens' Handbook of Experimental Psychology, 3rd Edition, Vol 4*. Wiley, New York. 1-41. Az Interneten 2016. szeptember 9-én: http://suppes-corpus.stanford.edu/articles/mpm/382.pdf

Narens, L. 1981. On the scales of measurement. *Journal of MathematicalPsychology, 24*, 249-275. Az Interneten 2016. szeptember 9-én: http://aris.ss.uci.edu/~lnarens/1981/Narens_JMP_1981.pdf

Roberts, F.S. 1979. *Measurement Theory with Applications to Decisionmaking, Utility and the Social Sciences.* Addison-Wesley, Maqssachusetts. Az Interneten 2016. szeptember 9-én: http://fitelson.org/roberts_measurement_theory.pdf

Steyer, R. 2001. Classical (psychometric) test theory. In Cook, T. és Ragin, C. (Szerk.) *International encyclopedia of the social and behavioral sciences. Logic of inquiry and research design*. Pergamon, Oxford. 1955 – 1962. Az Interneten 2016. szeptember 9-én: http://www.metheval.uni-jena.de/materialien/publikationen/ctt.pdf