Psihološka obzorja / Horizons of Psychology, 11, 3, 119-150 (2002)
© Društvo psihologov Slovenije 2002, ISSN 1318-187
Prevod znastvenega teoretsko-preglednega prispevka

Merjenje sprememb v skupinah in pri posameznikih s poudarkom na menjavi točk spremembe: Nova metodologija za ocenjevanje učinkov tretmaja z uporabo točk spremembe, osnovana na teoriji odgovora na postavko (TOP)#

Jörg A. Prieler[1] in John Raven[2]*
[1]Dr G. Schuhfried GmbH, Mödling, Avstrija
[2]Edinburgh, Škotska

Povzetek: V prvem delu tega članka opravimo pregled težav pri merjenju sprememb, predvsem pri izračunu spremembe doseženih točk (kot na primer točk, ki naj bi odražale medosebne razlike v odzivnosti na navodila, stres, terapijo ali zdravila). Drugi del tega članka opisuje nov pristop k diferencialnemu merjenju sprememb pri respondentih z različnimi sposobnostmi, predvsem izračun in uporabo točk spremembe.

Ključne besede: ocenajevanje, spremembe, Rasch, Flynnov učinek, Ravnove zahtevne matrice, selekcija kadrov

The measurement of change in groups and individuals with particular reference to the value of gain scores: A new IRT-based methodology for the assessment of treatment effects and utilising gain scores

Jörg A. Prieler[1] and John Raven[2]
[1]Dr G. Schuhfried GmbH, Mödling, Austria
[2]Edinburgh, Scotland

Abstract: Part I of this paper reviews problems in the measurement of change and, in particular, in the calculation of change scores (such as change scores purporting to index individual differences in such things as responsiveness to instruction, stress, therapy and drugs). Part II describes a new approach to the differential measurement of change among respondents having different levels of ability and, in particular, the calculation and use of change scores.

Prevod sta opravila David Gosar, abs. psih. in asist. Boštjan Bajec, pregledal pa izr. prof. dr. Valentin Bucik. Zahvaljujemo se avtorjema, ker sta dovolila uporabiti besedilo za prevod, priredbo in objavo v reviji Psihološka obzorja.

---

[*]Naslov / address: dr. John Raven, 30 Great King Street, Edinburgh EH3 6QH, Scotland, e-mail: jraven@ednet.co.uk

Key words: change, assessment, Rasch, Flynn effect, Raven's Advanced progressive matrices, personnel selection

CC=2260

Mnoge težave, vezane na ocenjevanje sprememb, še posebej v primeru posameznikov, ki na testih dosegajo visoke ali nizke rezultate, so poznane, kljub temu pa so v praksi pogosto zanemarjene. Te težave izhajajo iz učinkov stropa in neenakomernega povečevanja težavnosti testnih nalog na različnih delih lestvice. Poleg teh obstaja še veliko bolj resna težava, ki pa je manj poznana. Da bi v tem pregledu lahko opisali naravo le-te, moramo nekatere tehnične izraze uporabljati bolj ohlapno, tako da je njihov pomen nakazan prek njihovega konteksta. Bolj tehnično korekten opis težave se nahaja v opombi 1. Težava, docela opisana v pričujočem članku, je dejstvo, da se razlike v surovih dosežkih, ki odražajo enake razlike v latentnih sposobnostih, pomembno razlikujejo glede na absolutno težavnost testa, obliko njegove karakteristične krivulje in območje karakteristične krivulje testa, kjer so te razlike merjene. To drži tudi za teste, ki izpolnjujejo zahteve najpopularnejših različic modelov teorije odgovora na postavko (TOP - pogosto jih ohlapno imenujemo Raschev model). V preteklosti najpogosteje uporabljani pristopi prav zato ne omogočajo veljavnega podajanja ocene sprememb v primerih kot je podajanje ocene učinka specifičnega pedagoškega pristopa na relativni napredek bolj in manj sposobnih učencev. Iz tega sledi, da razlike v točkah, ki odražajo razliko v odzivnosti na nivoju *posameznika* (kot so točke »učnega potenciala« – sposobnost učenja iz navodil, točke »občutljivosti na šum«, ali točke »odzivnosti na določeno drogo«), odražajo različne stvari v različnih delih distribucije.

Drugi del članka opiše metodologijo, ki omogoča reševanje te težave. Čeprav je zasnovana na Raschevi teoriji, jo je mogoče uporabiti pri testih, ki niso skladni z Raschevim modelom. Zasnovana je na dveh domiselnih spoznanjih. Prvo predstavlja spoznanje, da uporaba testne naloge v dveh različnih trenutkih predstavlja "pomanjšano Raschevo lestvico", drugo pa je, da so spremembe v parametrih testnih nalog, ki sestavljajo Raschevo lestvico, mera sprememb pri *posameznikih*. Rezultat tega je izjemno prilagodljiv sklop ocenjevalnih postopkov za ocenjevanje sprememb s široko uporabnostjo.

Prvi del: Težave pri merjenju sprememb

Članek se osredotoča (a) na merjenje *diferencialnih* sprememb - npr. vpliva različnih tretmajev pri skupinah ljudi z različnimi sposobnostmi (ali so sposobnejši učenci z obogatitvenim programom pridobili več od manj sposobnih), in (b) na izračun in interpretacijo *individualnih* točk spremembe, kot v primeru merjenja "učnega potenciala" preko izračuna razlik med rezultati na *Ravnovih progresivnih matricah*,

pridobljenimi pred in po uvedbi treninga. V namen ilustracije problemov, ki kažejo na nujnost te razprave, in razvoja nove metodologije, ki bi te težave pri merjenju premostila, so v prispevku sprva predstavljene splošne težave pri merjenju sprememb.

Težave pri merjenju sprememb psiholoških karakteristik (o njih so razpravljali že avtorji, kot so Bereiter, 1963; Cronbach in Furby, 1970; Guthke, 1996; Harris, 1963; Lord in Novick, 1968; Rost, 1996; Williams in Zimmermenn, 1996) bodo v besedilu predstavljeni pod naslednjimi podnaslovi:

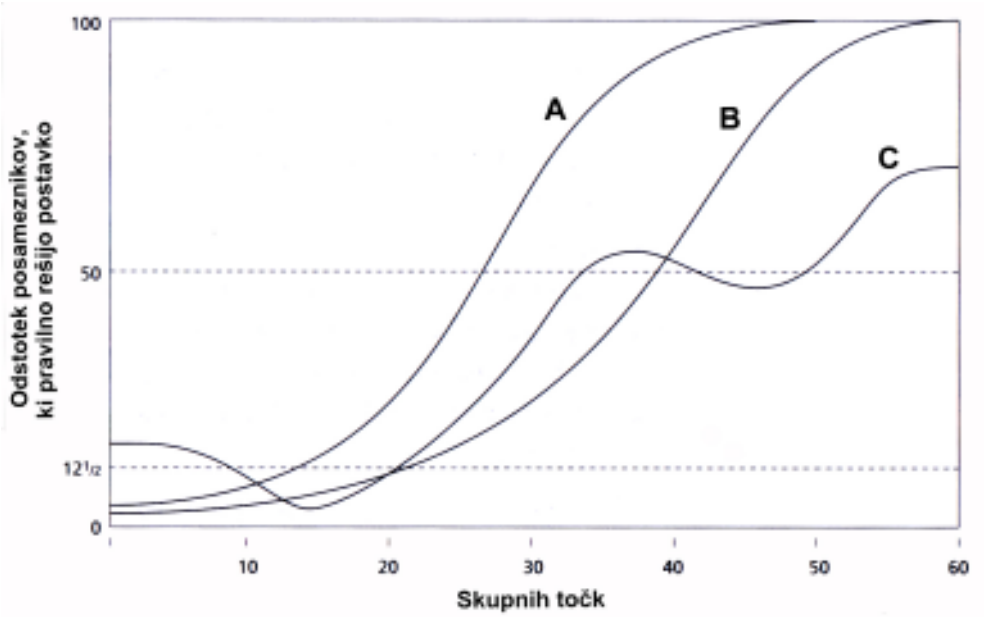
1. Težave, izhajajoče iz učinkov stropa in tal.
2. Težave, ki izhajajo iz pogoste potrebe po uporabi drugačnega in težjega testa po uvedbi intervencije (kot je npr. izobraževalni program), saj naj bi se znanje in sposobnosti *vseh* sodelujočih – kot posledica tretmaja - dramatično spremenile.
3. Težave, ki so pojavljajo, kadar obstoječi testi ne oblikujejo lestvic z ekvidistantnimi intervali. Tu se pojavljata dva sklopa težav:
   - težave, ki izhajajo iz neenakomerne probit distribucije testov, osnovanih na klasični testni teoriji in
   - težave, ki izhajajo iz dejstva, da razlike surovih rezultatov med visoko in nizko (za razliko od srednje) sposobnimi posameziki ne odražajo enakih razlik v latentnih sposobnostih; to drži tudi za teste osnovane na TOP in Raschevem lestvičenju oziroma skaliranju (glej spodaj); posledice tega so nepričakovane in pomembne, in ker je te probleme možno rešiti s pomočjo kasneje omenjene nove metodologije, bo večji del članka posvečen osvetljevanju tega specifičnega problema.
4. Težave, ki izhajajo iz preokupacije z uporabo mer, ki vključujejo le eno spremenljivko, ki pa izhajajo že iz težav pri uporabi velikega števila lestvic z več spremenljivkami; to so težave, ki izhajajo iz ne dovolj celostnih mer, ki jih prinaša preokupacija s klasičnimi merskimi lestvicami.
5. Težave, ki izhajajo iz nizke zanesljivosti, konstruktne veljavnosti in napovedne veljavnosti (in prek teh smiselnosti) razlik med individualnimi rezultati pred in po uvedbi tretmaja (kot sta trening ali stres) – točk prirastka oz. točk upada posameznika kot indicev globljih osebnih značilnosti (kot v primeru mer "učnega potenciala", "občutljivosti na stres" ali "moč reakcije na drogo").

V osnovnem besedilu tega prispevka smo se skušali omejiti na uporabo izrazov, ki - vsaj verjamemo tako - postajajo psihologom vse bolj znani. V pomoč tistim, ki se s temi termini še niso spoznali, skušamo razlago pojmov podati v opombah. Obenem podajamo tudi bolj formalne opise v končnih opombah. V odstavku, na katerega se nanaša ta opomba, se izraz "klasična testna teorija" uporablja za metodologijo, ki je najširše uporabljana pri izdelavi testov, ki jim je usojeno biti enodimenzionalni. Običajno to vključuje koreliranje postavk in faktorizacijo dobljene korelacijske matrike. Izraz "Teorija odgovora na postavko" (TOP) označuje namene razvoja testov, v katerih postavke tvorijo niz, pri katerem se posameznik strinja z oziroma pravilno reši vse postavke do določene točke, za njo pa se ne

strinja z oziroma ne reši pravilno nobene. Primeri takih testov so Guttmannove lestvice stališč, Ravnove zahtevne matrice in Britanske lestvice sposobnosti. Na področju fizike je primer tega uporaba metra za merjenje dolžine: kar je visoko 15 cm, je višje od 14, 13, 12… cm in nižje od 16, 17… cm. Bodite pozorni na to, da ne bi bilo smiselno vzpostavljati enodimenzionalnosti metra s tem, da bi korelirali in faktorizirali postavke.

Ustrezna metodologija za izdelavo podobnih lestvic na področju psihologije se je začela v Veliki Britaniji v sredini tridesetih let prejšnjega stoletja, v matematični obliki jo je v poznih štiridesetih letih prejšnjega stoletja zapisal Rasch na Danskem (Rasch, 1947), popularitzirali pa so jo Wright (1968) in drugi (Lord in Novick, 1968) v ZDA v zgodnjih šestdesetih letih prejšnjega stoletja. Teorija se je razvijala v koraku s tem (glej Birnbaum, 1968; Embertson, 1999; Fischer, 1974). Temeljna zahteva TOP je, da sestavljalci testov nekako pokažejo, da je naraščanje verjetnosti, da posameznik na postavko odgovori pritrdilno (oziroma pravilno) glede na njegovo lastnost (oziroma sposobnost) sistematično, in da je v sistematičnem odnosu z naraščanjem verjetnosti pri drugih postavkah. V sliki 1 liniji A in B odražata ta odnos, linija C pa ne (glej Hambleton, Swaminathan in Rogers, 1991 ali Raven, Raven in Court, 1998a za podrobnejšo razlago te slike).

V bolj formalnih obravnavah lastnosti ne izražamo v skupnem dosežku na testu, ampak z dosežkom "latnetne poteze". "Latentna poteza" je približno enaka latentnemu faktorju v klasični testni teoriji. Tehnično bolj ustrezno obravnavo teh tem bralec najde v opombi 2.



Slika 1: Hipotetično vedenje postavk na testu (povzeto iz Raven, Raven in Court, 1998a).

Težave, izhajajoče iz učinkov stropa in tal

Enostaven učinek stropa, ki se pojavi, ker sposobnejši udeleženci ne morejo pokazati vseh svojih sposobnosti, ker v testu ni dovolj dovolj težkih postavk, ali pa, ker za to, da bi vse svoje sposobnosti pokazali, ni dovolj časa, je dobro poznan. Manj pa so poznane posledice učinka stropa oziroma tal na zaključke znanstvenikov, ki so skušali opazovati spremembe kot posledice različnih obravnav (ne primer različnih učnih programov). Začnimo z enostavnimi primeri (ki bi bili banalni, če ne bi bili tako pogosti), kasneje pa bomo prikazali nekatere učinke, ki jih je težje zaznati. Ravno njihova zaznava predstavlja težavo, saj veliko znanstvenikov kljub priporočilu komisije Ameriškega psihološkega združenja (American Psychological Association, APA) za statistično pomembnost (APA, 1999), naj pred obdelavo pregledajo surove podatke, tega (zaradi očaranosti s sofisticiranostjo metod) ne naredi. S tem, ko povzamejo podatke s splošno sprejetimi metodami, je tudi zelo kritičnemu bralcu z dovolj časa in izkušnjami (kaj šele povprečnemu bralcu) težko opaziti večino "očitnih" učinkov stropa, ki jih bomo opisali.

Primer 1

Začnimo s preprostim, a prepogostim primerom, h kateremu se bomo še vrnili in ga kasneje natančneje pojasnili. Predpostavimo, da smo opredelili dve skupini managerjev - povprečno skupino in skupino "superzvezdnikov". Obe skupini sodelujeta v tritedenskem seminarju za razvoj managementa, oblikovanem za izboljšanje njihovih spretnosti. Lahko postavimo hipotezo, da bo povprečna skupina managerjev pridobila več kot skupina "superzvezdnikov", saj slednjim večina posredovanih spoznanj ni nova. Za testiranje hipoteze lahko oblikujemo test managerskih sposobnosti. Organiziramo testiranje vseh sodelujočih pred in po testiranju. Kasneje, kot večina drugih raziskovalcev, ugotovimo, da je napredek povprečno sposobnih managerjev večji od napredka bolj sposobnih. Naša hipoteza je potrjena.

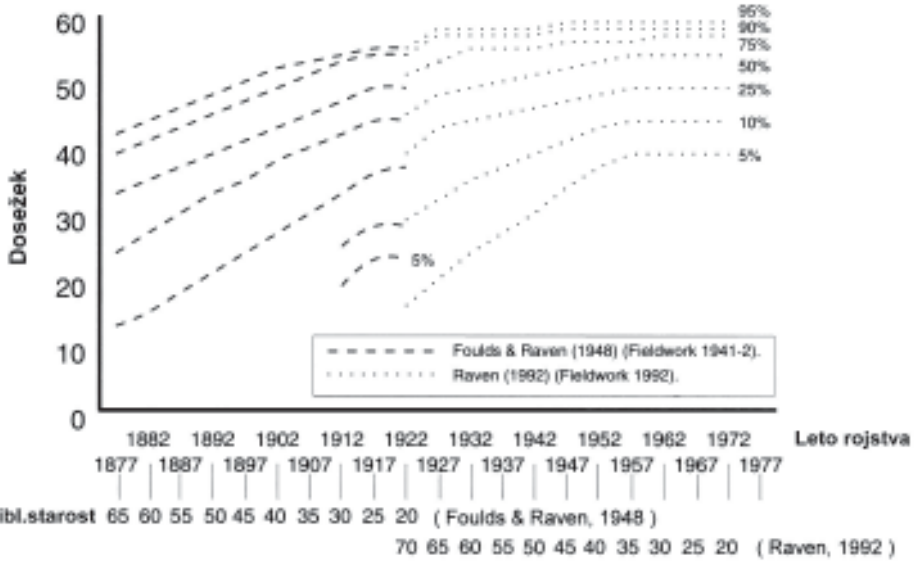
Kaj pa se je še lahko zgodilo? Kaj če je skupina "superzvezdnikov" že poznala večino odgovorov na testu že ob prvem testiranju, vendar je kljub temu pridobila mnogo novih znanj v teku seminarja. Svojega napredka pri drugem testiranju ne bi mogli pokazati. Lahko bi predpostavili, da to težavo enostavno prebrodimo, če podaljšamo test, vendar pa temu ni tako. Če bi enostavno dodali postavke, ki ustrezajo "superzvezdnikom", bi zlahka zamenjali svoj zaključek in dokazali, da so oni pridobili mnogo več od povprečnih managerjev. Na tej točki smo prišli do resnega metodološkega problema, ki je vplival na veliko število raziskav (kot so tiste, v katerih so skušali ugotoviti, ali z obogatitvenimi programi več pridobijo bolj ali manj sposobni študentje, ali tiste, v katerih so opazovali učinke različnih oblik poučevanja). Eden izmed ciljev pričujočega članka je predstavitev ozadja te težave in prikaz možnih rešitev.

Vendar predpostavimo nekaj drugega. Predpostavimo, da je to, kar so s

seminarjem pridobili "superzvezdniki", kvalitativno različno od tega, kar so pridobili povprečni managerji. Recimo, da so glavne pridobitve "superzvezdnikov" posledica interakcije z drugimi "superzvezdniki" in ne predavanj in primerov v seminarju. Tega ne bodo mogli izkazati na enodimenzionalnem testu managerskih znanj, vendar pa bi poskus, da bi uporabili niz nepovezanih postavk, da bi izsledili možne učinke seminarja, predstavljal hude težave za tradicionalne oblike redukcije podatkov in testiranja pomembnosti. K tej težavi se bomo vrnili pri podpoglavju 4, saj je to, da se ne uspemo osredotočiti na celostnost testiranja morda najbolj pomembna težava, s katero se trenutno spopadamo v psihometriji … To je težava, ki nam jo metodologija, ki jo imamo namen predstaviti, pomaga preseči.

Primer 2

Predstavili bomo še en primer metodoloških težav, ki jih predstavlja učinek stropa. Gre za primer, ki nam bo v končni fazi omogočil ilustracijo virov težav merjenja, ki jih želimo osvetliti, in tega, kako preseči te težave. Po zaslugi Flynnovih (1984, 1987, 1999) raziskav in objav je danes znano, da večina dosežkov na sestavljenih merah "splošne inteligentnosti" v zadnjem času dramatično narašča. Kot globalno oceno tempa in obsega naraščanja rezultatov Flynn navaja, da rezultati naraščajo za eno standardno deviacijo na generacijo. Učinek, izračunan preko standardnih deviacij na generacijo, je največji na merah rezoniranja (oz. pravilneje na "eduktivnih" sposobnostih)



Slika 2: 100 let eduktivne sposobnosti (povzeto iz Raven, Raven in Court, 1998a).

in najmanjši na merah znanja in rutinskih spretnosti - opredeljenih kot reproduktivne spretnosti. Naraščanje rezultatov mer eduktivnih sposobnosti, merjenih na podlagi Ravnovih zahtevnih matric, je razvidno s slike 2. Ta prikazuje percentilne norme pridobljene ob testiranju odraslih oseb različnih starosti (in torej različnih rojstnih datumov) s Standardnimi progresivnimi matricami (klasična oblika) v primeru testiranja okoli leta 1942 ter v primeru testiranja leta 1992. Približek starosti ljudi rojenih v različnih letih v obeh vzorcih je zapisana pod sliko.

Že ob prvem pogledu je razvidno, da so rezultati "manj sposobnih" - tistih, katerih rezultat se nahaja med petim in desetim percentilom - mnogo bolj narasli, kot rezultati bolj sposobnih - tistih, katerih rezultati se nahaja med 90. in 95. percentilom. Predstavljeni podatki prav tako jasno prikazujejo onemogočeno naraščanje rezultatov bolj sposobnih, kar je vsaj delno posledica učinka stropa na testu, ki s 60 postavkami tistim bolj sposobnim, rojenim v zadnjih letih, ne omogoča izkazovanja vseh svojih sposobnosti. Tako mnogi raziskovalci, ki preučujejo raziskovalne rezultate in zanemarjajo učinek stropa, zaključujejo, da rezultati manj sposobnih hitreje naraščajo kot rezultati bolj sposobnih.

Ko analiziramo raziskovalne rezultate težje oblike testa (Zahtevne progresivne matrice), ki so prikazani v tabeli 1, lahko takoj opazimo tudi drastični napredek bolj sposobnih. Žal pa še vedno ne vemo, če je učinek naraščanja večji ali manjši od učinka naraščanja pri manj sposobnih, saj smo uporabili drug test z drugačnimi merskimi karakteristikami. Niti rezultati, pridobljeni s testom Zahtevnih progresivnih matric, ne omogočajo konkretnejših zaključkov, saj je naraščanje tako veliko, da je tudi na tem testu mogoče zaslediti učinek stropa (test ima le 36 postavk).

K tej težavi se bomo vrnili kasneje. Velja opozoriti, da sta Teasdale in Owen (1989) preučila dosežke na latentnih potezah, ki sta jih pridobila s pomočjo testa, ki je od sodelujočih zahteval zaključevanje geometričnih oblik. Kljub temu, da v povzetku članka navajata da "nista odkrila dokazov za napredek na višjih nivojih", je na osnovi njunih rezultatov bolj korektno zaključiti, da sta odkrila splošen trend napredka tudi pri sposobnejših, vendar je bil trend napredka večji pri manj sposobnih. Študija je z vidika

Tabela 1. Zahtevne progresivne matrice, Sklop II: Primerjava med percentilnimi normami za odrasle v Veliki Britaniji 1962 in 1992.

	Starost (let)					
		20		30		40
Percentil	1962	**1992**	1962	**1992**	1962	**1992**
95	24	**33**	23	**33**	21	**32**
90	21	**31**	20	**31**	17	**30**
75	14	**27**	12	**27**	9	**26**
50	**9**	**22**	**7**	**22**	**--**	**20**

tega članka posebnega pomena, saj (i) je bila analiza izpeljana na točkah latetnih potez osnovanih na Raschevi analizi in (ii) je bila posebna pozornost namenjena preučevanju učinka stropa. Pred preučevanjem možnosti, da je le učinkoval prikrit učinek stropa, je potrebno primerjati konstruktno veljavnost testa, ki sta ga aplicirala Teasdale in Owen z Ravnovimi progresivnimi matricami, saj razlike v časovnih trendih med testoma lahko izvirajo iz različnih merskih konstruktov. Raven (2000) na podlagi študije literature, ki je vključevala študije Flynna (1999), Schaie in Willisa (1986), Thorndikea (1975) ter zaključuje, da mere eduktivne sposobnosti, merjene z verbalnimi in neverbalnimi testi, naraščajo približno za eno standardno deviacijo na generacijo, mere reproduktivnih sposobnosti pa se skorajda ne spreminjajo, medtem ko mere, pridobljene s pomočjo testov, ki so osnovani na obeh komponentah kognitivnih sposobnosti, naraščajo v skladu s prevladujočo faktorsko zasičenostjo z obema osnovnima komponentama. Ta zaključek dodatno podpira Flynnova (2000) študija podlestvic Wechslerjevega testa.

Test, ki sta ga uporabila Teasdale in Owen, kaže naraščanje 0.5 standardne deviacije na generacijo, kar daje sklepati, da je vseboval tako eduktivne kot tudi reproduktivne sposobnosti - v takem primeru gre, na podlagi prehodnih raziskav, pričakovati manj jasne rezultate. Ali obstaja možnost prikritega učinka stropa? Seveda obstaja, saj je bil test – poleg tega, da je šlo za test moči – časovno omejen, kar je lahko onemogočilo sodelujočim, da bi v celoti izkazali svoje sposobnosti.

Ob merjenju sprememb je pogosta napaka prav zamenjevanje hitrosti in moči. Obstajata dva razloga za to:

- Zaradi časovne omejitve mnogi sodelujoči ne rešijo postavk na koncu testa, kar onemogoča izračun pravih parametrov postavk za le-te, vključno s stopnjo težavnosti.
- »Učinek stropa kot posledica časovne omejitve« neposredno izhaja iz časa, potrebnega za rešitev postavke. To je najlaže razvidno iz primera. Raziskovalec uporabi test s 60 zelo lahkimi postavkami in petminutno časovno omejitvijo za reševanje. Predpostavimo, da je izjemno sposobna oseba komaj zmožna v tem času pravilno rešiti vseh 60 postavk. Sodelujoči se nato udeleži treninga, ki pomembno izboljša sposobnosti vseh v skupini. Raziskovalec, z namenom, da bi meril spremembe, podaljša test za 20 postavk. Kljub temu, da je sodelujoči mnogo bolj izurjen, kot je bil, še vedno potrebuje za odgovor na postavko pet sekund. Njegov rezultat še vedno znaša 60 točk!

Težave, ki izhajajo iz dejstva, da je po intervenciji pogosto potrebno uporabiti drugačen in težji test

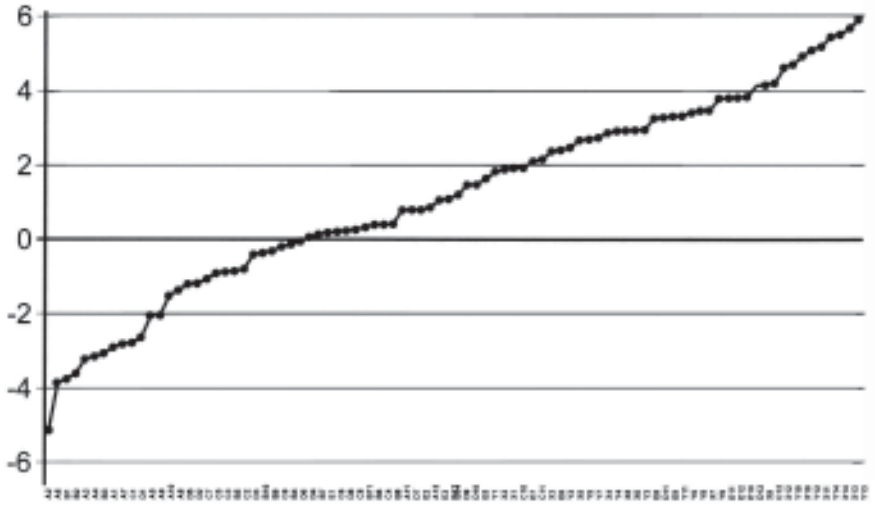
S temi težavami smo se srečali že dvakrat: prvič, ko smo opazili, da program treninga za managerje lahko izboljša sposobnosti bolj sposobnih preko stopnje, ki jo test lahko še meri, drugič pa v primeru dokumentiranja relativnih sprememb v rezultatih manj in

bolj sposobnih skupin v populaciji na Ravnovem testu progresivnih matric. Ocena relativnih sprememb je problematična, saj test nezadovoljivo diskriminira pri manj sposobnih, primerno pri srednje sposobnih, in ponovno nezadovoljivo pri zelo sposobnih. Kljub dejstvu, da so bili podatki zbrani s težjim testom in so razkrili napredek tudi pri visokih rezultatih, jih ni bilo mogoče pretvoriti v primerno mersko enoto, ki bi nudila odgovor na vprašanje, ali je napredek v določenem smislu uniformen preko vseh ravni sposobnosti.

To so le konkretni primeri zelo splošnega problema dokumentiranja dolgoročnih vplivov programov obogatitve izobraževalnih programov, znotraj katerih je potrebno uporabiti teste zelo različnih težavnosti za testiranje skupin posameznikov različnih starosti (naj omenimo, da je to praktična težava, ki jo je Rasch (1947, 1960/1980) srečal pri opazovanju dolgoročnih učinkov eksperimentalnega programa branja, ko je moral (nujno) uporabiti različne teste pri različnih starostih, ko so učenci napredovali v teku šolanja, in ga je spodbudila k izoblikovanju matematične različice TOP).

Težave, ki so pojavljajo, kadar obstoječi testi ne oblikujejo lestvic z ekvidistantnimi intervali in težave, ki izhajajo iz neenakomerne probit distribucije testov, osnovanih na klasični testni teoriji

Te težave lahko ilustriramo s pomočjo slike 3. Ta prikazuje težavnost 84 postavk, izraženo v Raschevih logit enotah, oblikovanih v teku razvoja testa (test Standardne progresivne matrice Plus), ki naj bi imel enako diskriminativno moč v zgornjem razponu



Slika 3: Standardne progresivne matrice Plus – težavnosti postavk (logit): 84 poskusnih postavk iz izenačitvene študije postavk leta 1996 (povzeto iz Raven, Raven in Court, 2000).

sposobnosti, kot jo je imel Ravnov test Standardne progresivne matrice, ko je bil izdan prvič. Slednji je, kot je razvidno iz slike 2, zaradi porasta skozi generacije izgubil diskriminativno moč v zgornjem razponu sposobnosti.

Če se ozremo na sliko 3, lahko opazimo več platojev. Na teh točkah se surovi rezultati posameznikov povečajo za eno točko za vsako testno postavko na katero odgovorijo pravilno. Razvidno pa je tudi, da so razlike, na katere kažejo nivoji latentne sposobnosti, dobljene na osnovi surovih rezultatov, minimalne. Carver (1989) opozarja, da je to vodilo k nekaterim zavajajočim raziskovalnim zaključkom. Na primer, ko posamezni otroci odraščajo in dosežejo enega izmed teh platojev, se njihov surovi rezultat nenadno poveča. To je prispevalo k pojmovanju intelektualnega razvoja kot stopenjsko kontinuiranega - obstajajo obdobja, ko se rezultati nenadno povečajo in obdobja, ko ostajajo stabilni. Domneve o tem, koliko postavk v določenem delu lestvice je "preveč" ali "premalo", je zelo odvisna od sposobnosti in starosti vključenih v raziskavo, saj je očitno, da so surovi rezultati razlik med posamezniki in v času močno odvisni od naklona grafa v delu, na katerega se osredotočamo. To drži ne glede na pogoste pritožbe raziskovalcev, ki se pri svojem delu osredotočajo na občutljivost ljudi določene starosti ali izraženosti sposobnosti na učinke tretmaja, da je v testih »preveč« nepomembnih postavk in »premalo« postavk na določenem delu lestvice.

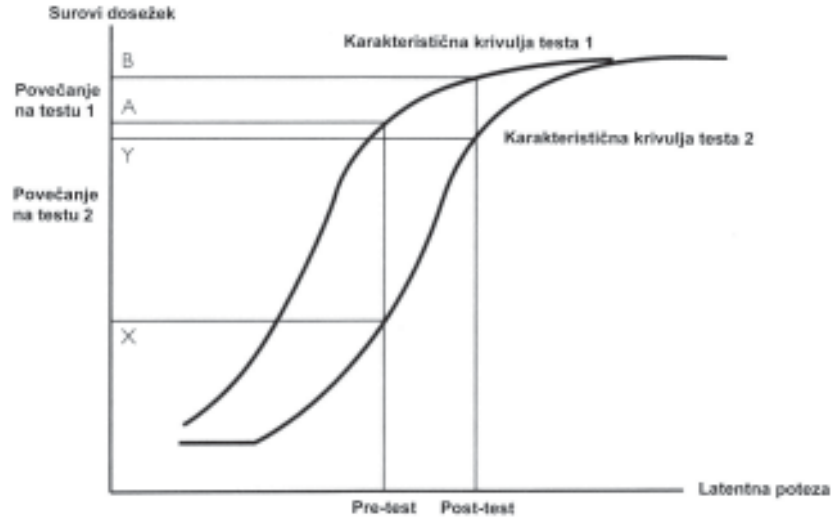
Na tej točki se je poučno, z ozirom na uporabo statistik, namenjenih povzemanju rezultatov, vrniti k sliki 2 in preučiti Flynnovo trditev, da znaša prirastek točk na SPM eno standardno deviacijo na generacijo. Kot lahko vidimo, je 50. percentil (ki bi v primeru Gaussove distribucije bil enak aritmetični sredini) za posameznike, rojene leta 1877, znašal 24 točk. Za posameznike, rojene leta 1972, je znašal 54 točk. Torej znaša sprememba medianskih vrednosti v stoletju, ki ga prikazuje slika, 30 točk. Vendar pa je predstavitev rezultatov v obliki standardnih deviacij problematična. Kot je razvidno iz slike, distribucije niso normalne in se razlikujejo glede na leto rojstva. Vsak študent psihologije zna izračunati standardno deviacijo s pomočjo SPSS ali kakega drugega programskega paketa. In vsak študent psihologije ve (ali pa je vedel), da je 68% dosežkov znotraj obsega aritmetične sredine ± ene standardne deviacije. Iz tega sledi, da lahko standardno deviacijo katerihkoli podatkov odčitamo iz prikaza, kakršnega imamo na sliki 2. To lahko naredimo tako, da ocenimo obseg rezultatov, ki zajema 34% populacije nad ali pod aritmetično sredino. Kot je razvidno iz slike, predstavlja ta del, za posameznike rojene med 1900 in 1930, 12 točk, če podamo oceno na podlagi rezultatov pod mediano, in 8 točk, če podamo oceno na podlagi rezultatov nad mediano. Torej bi lahko prirastek 30 točk na 30 let lahko grobo ocenili kot nekaj manj kot eno standardno deviacijo za obdobja, ki jih obsegajo podatki. Vendar pa v primeru, da podamo oceno standardne deviacije na osnovi rezultatov najmlajše starostne skupine na osnovi razpona nad mediano, pridemo do ocene treh točk - prirastek v letih, ki jih prikazuje slika, tako obsega 10 standardnih deviacij. To seveda ni konec zgodbe, saj je testni učinek stropa zmanjšal tako mediano kot tudi varianco dosežkov najmlajše starostne skupine. Ekstrapolacija krivulj - ki je, kot vemo, preko našega dela z Zahtevnimi progresivnimi matricami, upravičena - poda oceno prave mediane, za tiste, ki so

rojeni leta 1972 in testirani leta 1992, ki znaša 70 točk. Torej znaša "resnični" prirastek medianskega rezultata, v obdobju, ki jih obsegajo rezultati, 46 in ne 30 točk. "Resnična" standardna deviacija znaša torej 10 točk. "Resnični" prirastek znaša 1.5 standardne deviacije za 3.1 genaracij vsebovanih v podatkih, če znaša ena generacija 30 let. V primeru, da ena generacija znaša 20 let, znaša prirastek .9 standardne deviacije. Iz zapisanega sledi, da poskusi popravka odstopanj distribucij surovih rezultatov preko pretvorbe v standardne deviacije, ki se jih učimo v običajnih študijskih programih statistike in jih zahteva večina urednikov revij, zelo verjetno pripeljejo do zavajajočih ocen velikosti sprememb v času.

Težave, ki izhajajo iz dejstva, da razlike surovih rezultatov med visoko in nizko sposobnimi posamezniki ne odražajo enakih razlik v latentnih sposobnostih.

Kot uvod v predstavitev teh težav se lahko ponovno vrnemo na prošnjo našega hipotetičnega sponzorja, da ocenimo relativen napredek povprečnih managerjev in managerjev "superzvezdnikov", ki je posledica izvedenega izobraževanja. V naslednjem koraku bomo, presenetljivo, pokazali, da težave ne moremo rešiti ne z razvojem težavnejšega testa in ne z razvojem testa, skladnega z Raschevim modelom.

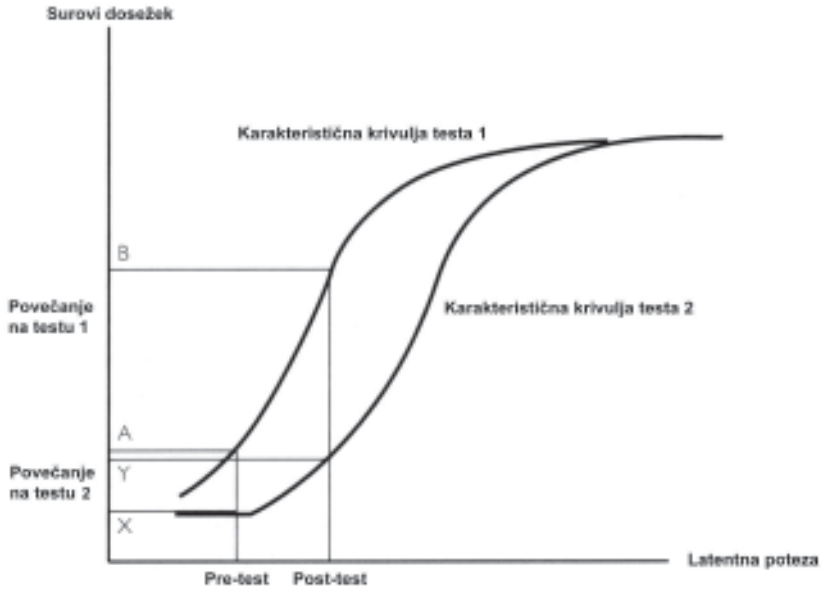
Slika 4 prikazuje to težavo za visoko sposobne kadre, slika 5 pa za nizko sposobne kadre. Če uporabimo test z njegovo karakteristično krivuljo, prikazano v levem delu slike 4, se aritmetična sredina visoko sposobne skupine zveča od A na prvem testiranju



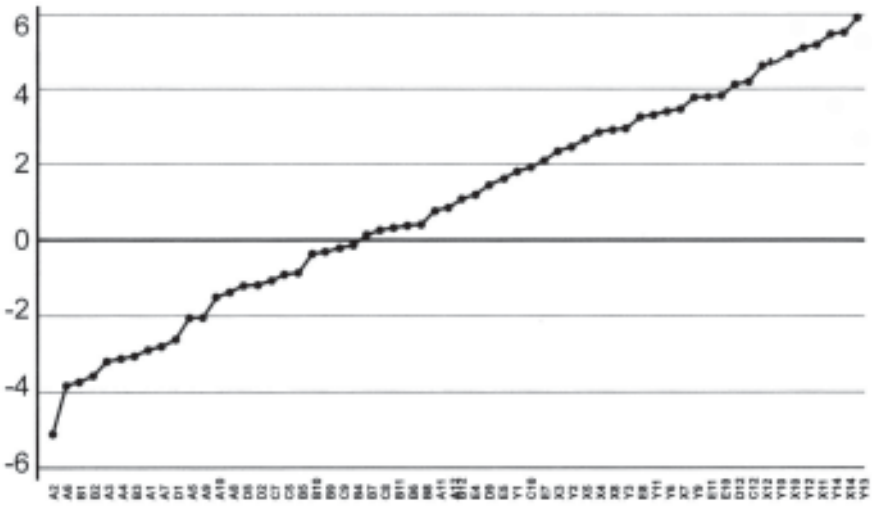
Slika 4: Ilustracija sprememb v surovih dosežkih na »lahkem« in »težkem« testu »managerskih sposobnosti« za enake spremembe v latentni sposobnosti. Prikaz za zelo sposobne (povzeto iz Raven, Raven in Court, 2000).

(pred treningom) do B na drugem testiranju (po treningu). Razlika predstavlja razmeroma majhen prirastek. V primeru uporabe težjega testa, prikazanega v desnem delu slike, se enak prirastek v latentni sposobnosti visoko sposobne skupine odraža v ogromnem prirastku v surovih rezultatih (od X do Y). Slika 5 prikazuje prav nasprotno situacijo pri nizko sposobnih. Navidezni prirastek, opredeljen kot razlika med prvim in drugim testiranjem, je ogromen na testu 1 in zanemarljiv na testu 2.

Sinteza obeh primerov nam pokaže, da v primeru, ko raziskovalec uporabi test 1 za merjenje napredka obeh skupin, odkrije velik relativni napredek nizko sposobne skupine in majhen napredek visoko sposobne skupine. Ob uporabi testa 2 raziskovalec odkrije prav nasprotno. Splošen in izrazito pomemben zaključek, ki sledi iz navedenih primerov, je, da je navidezna velikost prirastka v latetnih sposobnostih, ki izhaja bodisi iz razvojnih izkušenj, bodisi iz naravnih sprememb, odvisna od (a) splošne težavnosti testa, ki je relativen glede na sposobnosti sodelujočih, ter (b) distribucije parametrov postavk, ki so relativne glede na interval v latetni sposobnosti, kjer sprememba poteka. Brez uporabe pristopov, katerih predstavitev sledi v nadaljevanju članka, je zato praktično nemogoče podati kakršnekoli smiselne ocene relativnega napredka visoko, srednje in nizko sposobnih skupin. Bolj konkretno, iz navednih primerov in splošnih zaključkov sledi, da ne moremo dokumentirati relativnega napredka visoko in povprečno sposobnih managerjev z izločanjem lažjih in dodajanjem težjih postavk. Prav to bi



Slika 5: Ilustracija sprememb v surovih dosežkih na »lahkem« in »težkem« testu »managerskih sposobnosti« za enake spremembe v latentni sposobnosti. Prikaz za nizko sposobne (povzeto iz Raven, Raven in Court, 2000).



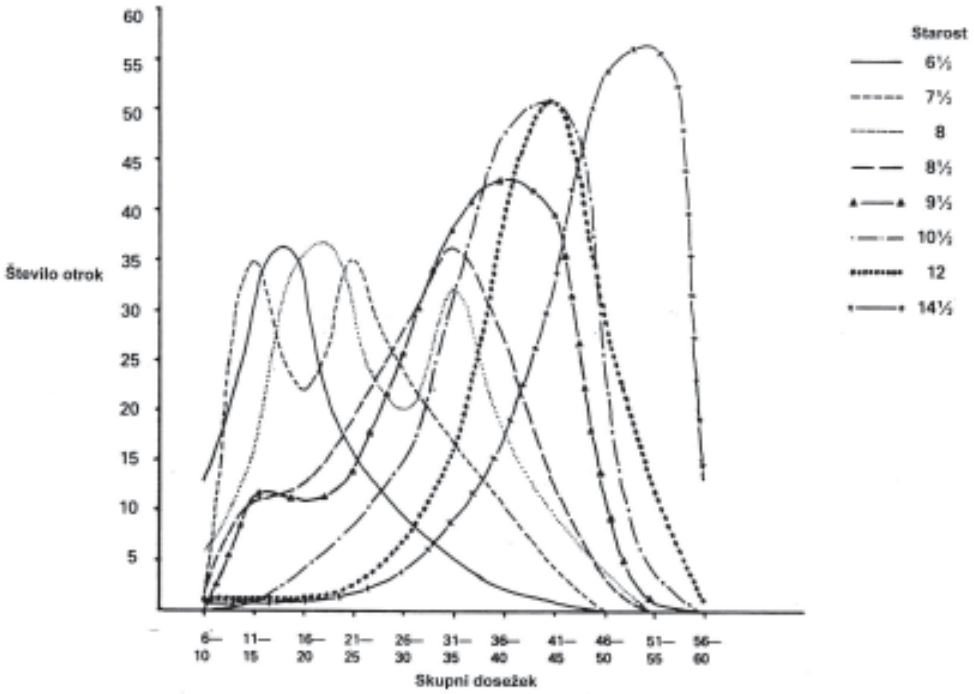
Slika 6: Standardne progresivne matrice Plus – težavnosti postavk (logit): končnih 60 postavk testa iz izenačitvene študije leta 1996 (povzeto iz Raven, Raven in Court, 2000).

preoblikovalo karakteristično krivuljo testa iz krivulje, prikazane na levi, v krivuljo, prikazano na desni. Kaj pa, če se bi se poslužili testa z linearno karakteristično krivuljo testa? Platoji na grafu, prikazanem v sliki 3, so vodili k izločitvi nekaterih testnih postavk. To je vodilo k izboru 60 testnih postavk, katerih težavnosti so prikazane na sliki 6.

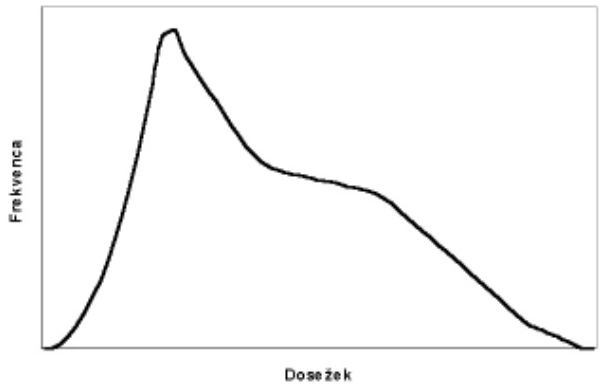
Na prvi pogled nam to nudi rešitev naše težave - pod pogojem, da bi tak test uporabili za obe testiranji (prvo in naknadno testiranje). Na žalost skladnost z Raschevim modelom ne zagotavlja lestvice z ekvidistantnimi intervali. Prav tako ni nobenega zagotovila, da bo prikazana splošna distribucija testa, kakršne je, podobna distribucijam različno sposobnih skupin. Resnost te težave lahko prikažemo z drugimi resničnimi podatki. Slika 7 prikazuje distribucije klasičnega SPM testa znotraj posameznih starostnih skupin (ne testa SPM Plus, katerega podatki, pridobljeni v teku razvoja, so bili že vključeni v diskusiji).

Takoj je razvidno, da so te distribucije znotraj posameznih starostnih skupin, kot tudi distribucije posameznih podlestvic pri posameznih starostnih skupinah na mnogih večkomponentnih testih "inteligentnosti" bimodalne in vse prej kot normalne. Kot sledi iz osebne komunikacije z Robertom Thorndikeom, velja to tudi pri distribucijah znotraj starostnih skupin znotraj posamezne podlestvice pri veliko večkomponentnih testih »inteligentnosti«. Kaj pa če jih seštejemo? Ali ne bi po tem dobili Gaussove distribucije? Distribucija seštevkov rezultatov je prikazana na sliki 8.

Glede na navedeni primer si je težko predstavljati, kako je splošna distribucija rezultatov slehernega testa, razvitega v skladu z Raschevimi zahtevami po normalnosti



Slika 7: Standardne progresivne matrice (klasična oblika): distribucija testnih dosežkov za osem starostnih skupin (povzeto iz Raven, 1981).



Slika 8: Standardne progresivne matrice (klasična oblika): skupna distribucija testnih dosežkov za 3466 otrok, starih od 6 do 15 let (povzeto iz Raven, 1989).

znotraj starostnih distribucij, lahko normalna (kot je potrebno predpostaviti v primeru vseh do sedaj omenjenih poskusov reševanja težav).

Na tej točki se je smiselno vrniti k opažanjem Teasdalea in Owna. Kljub temu,

da v svojem besedilu opozarjata na dejstvo, da so se rezultati visoko sposobnih sodelujočih komajda povečali, je iz slike 1 v njunem članku očitno, da so se tudi rezultati najmanj sposobne skupine sodelujočih zelo malo spremenili. Ne nasprotujemo njuni trditvi, da ni bilo prirastka surovih rezultatov ekvivaletnih 95. percentilu, in da so prisotne pomembne spremembe surovih rezultatov ekvivaletnih 5. percentilu. Želimo opozoriti na dejstvo, da je razlog za navidezno odsotnost sprememb tudi najmanj sposobnih zgolj posledica karakteristične krivulje uporabljenega testa (poleg tega velja omeniti, da so se standardne deviacije testa, ki sta ga uporabila Teasdale in Owen, pomembno znižale med prvim in drugim testiranjem). Karakteristična krivulja testa je mnogo bolj strma v primeru drugega testiranja. Izguba diskriminativne moči, v zadnjem času značilna tudi za RPM, ni razvidna iz rezultatov, v katerih Bouvier (1969) poroča o izsledkih uporabe mnogih testov na Belgijskih rekrutih med leti 1958 in 1976.

Težave, ki izhajajo iz preokupacije z uporabo mer, ki vključujejo le eno spremenljivko, ki pa izhajajo že iz težav pri uporabi velikega števila lestvic z več spremenljivkami; težave, ki izhajajo iz ne dovolj celostnih mer

Na tem mestu se spet lahko vrnemo k naši primerjavi pridobitev managerjev »superzvezdnikov« in povprečnih managerjev. Prej smo omenili, da so lahko te pridobitve kvalitativno različne in se zato ne bi pokazale na enodimenzionalnem testu managerskega znanja. Uporaba ne dovolj celostnih instrumentov za evalvacijo, ki so neprimerno prilagojeni ciljem in sredstvom učnih programov, da bi le-te primerjali med seboj po njihovih želenih in neželenih učinkih na različne učence, je pripeljalo do kopice primerjalnih evalvacijskih študij, ki niso le nekompetentne, ampak tudi neetične (Raven, 1991, 2000a; Raven in Stephenson, 2001). Naj na kratko navedemo primer – številne evalvacije »odprtih« ali »naprednih« učnih programov so pokazale, da ne izboljšajo bralnih, pisnih ali matematičnih spretnosti udeležencev, kadar jih merimo konvencionalno. Vendar pa glavni cilji večine programov »odprtega« ali »progresivnega« poučevanja niso na tem področju, ampak na področju izboljšanja samozavesti, sposobnosti komuniciranja, dela z drugimi, na področju promocije razvoja različnosti – različnih talentov pri različnih otrocih. Še več – glavne slabosti tradicionalnih oblik poučevanja so ravno v razbijanju pozitivne samopodobe, vzgajanju občutkov naučene nesposobnosti in v njihovem ustvarjanju monokultur uma namesto različnosti. Nobene primerjalne študije, ki ne razišče tovrstnih potencialnih prednosti ali slabosti, ne moremo obravnavati kot kompetentno ali objektivno znanost. Vendar so tovrstne neustrezne evalvacije – katerih največja pomanjkljivost je pomanjkanje celostnosti – pripeljale do ukinitve programov. Tu ne gre le za škodljiv učinek na učence, ki imajo potencial, da do visoke stopnje razvijejo različne spretnosti, ki bi jih odprti ali progresivni programi morda lahko negovali. Gre za to, da so ukinjani programi edini programi, ki negujejo spretnosti, ki jih bodo bivši učenci potrebovali, če naj bi spremenili našo družbo tako, da bo imela naša vrsta možnost preživetja (Raven, 1994, 1995; Raven in Stephenson, 2001). Težko si je zamisliti kaj bolj neetičnega.

Izkaže se, da je temelj znanstvene objektivnosti celostnost testiranja, ne pa natančnost testiranja na posamezni spremenljivki. Eden dejavnikov, zaradi katerih

tega v preteklosti nismo sprevideli, je bila preokupacija z enodimenzionalnimi lestvicami z več postavkami. Kako lahko posameznik v realni študiji doseže celostnost, ki bo zadovoljila zahteve stroke po testiranju statistične pomembnosti, ne da bi sestavil ogromno baterijo testov (ki jih je potrebno še izdelati), ki pa nam še vedno ne bi odgovorili na vprašanje, ali so se posamezni učenci razvili v lastnih, vendar življenjsko pomembnih smereh. Nič ne bi moglo bolje predočiti potrebe po dveh stvareh. Prva je razvoj, ki ga bomo opisali kasneje. Druga je psihometrični model, katerega opis presega namen tega članka, je pa opisan v Raven in Stephenson (2001) in podobnih publikacijah (za bralca bo morda zanimivo, da je na to težavo opozoril že Spearman 1926-ega leta, ko je zapisal »Vsak normalen moški, ženska in otrok… je genij v nečem… Odkriti je treba v čem… To je zagotovo najtežja naloga, saj se pojavlja le v majhnem deležu vseh možnih sposobnosti. Zagotovo se tega ne da zapaziti z nobenim izmed načinov testiranj, ki so trenutno v uporabi. Vendar pa verjamem, da je te načine možno zelo izboljšati«).

Težave, ki izhajajo iz nizke zanesljivosti, konstruktne veljavnosti in napovedne veljavnosti (in prek teh smiselnosti), razlik med individualnimi rezultati pred in po uvedbi tretmaja (kot sta trening ali stres) – individualnih točk prirastka oz. točk upada kot indicev globljih osebnih karakteristik (kot v primeru mer "učnega potenciala", "občutljivosti na stres" ali "moč reakcije na drogo")

Na tem mestu dobo obravnavani trije sklopi težav: (i) tiste, ki izhajajo iz problemov nizke zanesljivosti individualnih sprememb rezultatov, (ii) tiste, ki izhajajo iz dejstva, da tudi testi, sestavljeni na podlagi teorije odgovora na postavko in testi z zanemarljivim testnim učinkom stropa izkazujejo negativno korelacijo prvotnega testnega rezultata s kasnejšo spremembo rezultata in (iii) tiste, ki izhajajo iz dejstva, da imajo lahko razlike, sicer enakovredne v smislu konstruktne veljavnosti preučevanih mer, drugačen pomen in pomembnost na različnih točkah lestvice (Bereiter, 1963) in s tem neuniformno interpretacijo ali konstruktno veljavnost točk prirastka (npr.: posledični prirastek petih točk po uvedbi Feursteinovega programa instrumentalne obogatitve pri nekom, katerega rezultat na RPM je zelo nizek, ima precej drugačen pomen in pomembno drugačne posledice, kot identičen rezultat prirastka nekoga z zelo visokim rezultatom.)

Bralcu bo morda lažje razumeti omenjeno problematiko, če se osredotočimo na najbolj odmevno uporabo točk »prirastka«. Mnogi avtorji (npr. Budhoff, 1973; Budhoff, Corman in Gimon, 1976; Guthke, 1982; Guthke in Wiedl, 1996), še posebej Feurestein (1979) in Feuerstein, Klein in Tannenbaum (1990) so domnevali, da naj bi sposobnost napredka v izobraževalnem programu, osnovanem na poučevanju sodelujočih, kako reševati probleme, imela večjo napovedno veljavnost kot rezultati testa mer "sposobnosti reševanja problemov" pred izvedbo izobraževalnega programa. Ta sposobnost, pogosto opredeljena kot "učni potencial", je bila večinoma merjena preko spremembe v RPM rezultatih pred in po izobraževalnem programu kot je Feursteinov program »instrumentalne obogatitve«. Posameznikovi rezultati pred programom se odštejejo

od rezultatov po programu in tako oblikujejo rezultat "prirastka". Slednji služijo kot mera učne sposobnosti oz. "učnega potenciala".

Govorili bomo o smiselnosti takih individualnih rezultatov. Kasneje bomo razpravljali o možnostih, kako preseči te težave. Naprej pa bi želeli povezati to, o čemer bomo spregovorili, s težavami, o katerih smo razpravljali poprej, saj ni samo pomen in vrednost mer "učnega potenciala" edino, kar moramo preučiti za razumevanja dela teh avtorjev. Raziskovalci pogosto trdijo, da otroci z "nizkimi rezultati" oz. "prikrajšani" otroci pridobijo več od učnih programov kot bolj sposobni. Videli smo že, kako težko je utemeljiti take trditve.

i *Težave, ki izhajajo iz nizke zanesljivosti individualni sprememb rezultatov.*
Ocenjevanje vloge napake v točkah sprememb je paradoksalna tema v klasični testni teoriji. Ko korelacija med merami pred in po obravnavi upada, upada relativna napaka v točkah sprememb (glej Lord, 1963 za klasičen obrazec za zanesljivost točk sprememb). Vendar pa, kot je izpostavil Breiter (1963), nizka korelacija mer pred in po obravnavi pripelje do težav v interpretaciji pomena »spremembe«, saj kaže na to, da testi ne merijo iste dimenzije. Paradoks je posledica tega, da ne konceptualiziramo spremembe kot ločene dimenzije (Embretson, 1991).

ii *Težave, ki izhajajo iz dejstva, da tudi testi, sestavljeni na podlagi teorije odgovora na postavko, in testi z zanemarljivim testnim učinkom stropa, izkazujejo negativno korelacijo prvotnega testnega rezultata s kasnejšo spremembo rezultata.*
Naslednji klasični paradoks je, da ima korelacija med točkami spremembe in začetno sposobnostjo negativno pristranost zaradi skupnih varianc napake. Komponenta napake v prvi sposobnosti je vključena tudi v točke spremembe, vendar pa s spremembo predznaka (Lord, 1963). Tako pride do negativne pristranosti v korelaciji točk spremembe z začetno stopnjo (Embertson, 1991). Ta pristranost se lahko zmanjša z maksimizacijo natančnosti merjenja na vsaki stopnji; na primer z uporabo testov TOP, prilagojenim testiranjem, uporabimo lahko dve neodvisni meri z istim testom za definicijo točke 1 v času (prvo mero uporabimo za opis začetne sposobnosti, drugo pa za izračun točk prirastka do naslednje točke v času – Fischer, 1974).

iii *Težave, ki izhajajo iz dejstva, da imajo lahko razlike, sicer enakovredne v smislu konstruktne veljavnosti preučevanih mer, drugačen pomen in pomembnost na različnih točkah lestvice, kar lahko vodi k variabilnosti konstruktne veljavnosti točk prirastka.*
Ta problem lahko ilustriramo na študiji, ki jo je naročila Avstrijska vojska (Prieler, 1998, 2000). Njen cilj je bil oceniti relativno veljavnost širokega razpona različnih individualnih sprememb - oz. razlike - rezultatov, ki so odražali reakcije na stres

in so služili kot prediktor kasnejšega uspeha. Nekatere izmed teh so bile "prirastki", druge pa "upadi". Ob začetku programa je skoraj 3000 oficirjev izpolnilo testno baterijo psiholoških testov. Po stresogenem dogodku (oficirjem ni bilo dopuščeno spati in so bili primorani hoditi skorajda celo noč) so bili testirani ponovno. Raziskovalno vprašanje je bilo, ali bodo testni rezultati prvega testiranja ali pa rezultati razlik boljši prediktorji uspeha v sledečem 12 mesečnem vojaškem treningu. V primeru višje napovedne vrednosti rezultatov razlik je avtorja zanimalo tudi, kateri rezultati razlik so bili najboljši prediktorji. Večino težav, ki jih je vsebovala dosedanja razprava, je vključenih v tej študiji: na rezultate razlike (prirastek ali upad) pomembno vplivajo relativne standardne deviacije različnih testov, položaj sodelujočih v testni distribuciji, izravnanost probit distribucije, testni učinek stropa in aplikacija težavnejših testov po indukciji stresogenega dogodka. Iz tega sledi, da zgolj osnovnih rezultatov razlik ne moremo uporabiti na smiseln način.

Drugi del: Pot naprej - metodologija, ki jo je uporabil Fischer

Po tem, ko smo prikazali, da je opazovanje sprememb, še posebej diferencialnih sprememb na različnih stopnjah sposobnosti, prežeto s težavami, se sedaj osredotočimo na tisto, kar lahko storimo. Opisano metodologijo je razvil Fischer na Dunaju (Fischer, 1972, 1974, 1983, 1995) kot odgovor na članke avtorjev kot so Bereiter (1963), Cronbach in Furnby (1970) in Holzmann (1963). Čeprav so nekateri izmed omenjenih avtorjev raziskovali predvsem razlike med skupinami, drugi (na primer Klauder, 1991, Liou, 1993, Ponocny, 2000) pa so delali na spremembah pri posamezniku, se osredotočamo na Fischerjevo delo, saj menimo, da je najbolj prilagodljivo, posplošljivo in splošno uporabno.

Kmalu po tem, ko je Bereiter v letu 1963 izdal svoj članek, je Fischer začel iskati rešitev problema merjenja sprememb, v glavnem za opazovanje sprememb skupin. 1972 je izdal svoj prvi članek o merjenju sprememb v skupinah. Po tem je izšlo več njegovih prispevkov v obliki poglavij knjig (na primer 1974, 1977, 1995, 1997) in člankov (na primer 1983, 1989, 1995). Čeprav je bilo njegovo delo široko sprejeto, pa njegova spoznanja niso bila uporabljana. Glavni razlog temu je bilo izračunavanje parametrov učinka, ki je bilo brez uporabe osebnega računalnika zelo zapleteno. Leta 1974 je Fischer izdal izvorno kodo, ki jo je bilo potrebno prevesti in izvesti na velikih računalnikih. Uporabniku ni bila prijazna in tako so jo uporabljali le njegovi študentje. V 1980 je postala dostopna DOS različica. Bila je nekoliko bolj prijazna uporabniku, vendar pa dostopna le študentom na univerzi na Dunaju. V tem času osebni računalniki še vedno niso bili posebej dostopni. Fischer je na univerzi vseeno poučeval uporabo tega programa. Uporabniško prijazna programska oprema za opazovanje sprememb skupine in posameznika je bila izdelana 1998. Distribuira jo uraden distributer programske opreme in ni več dostopna le študentom.

Obstajata dve glavni področji uporabe metodologije, ki je bila razvita:

1. Merjenje in statistično ocenjevanje sprememb v skupinah (a) v teku časa, (b) kot posledic različne vrste in obsega tretmajev (c), kot posledic istih tretmajev, ne glede na nivoje sposobnosti, in (d) v skupinah, ki se razlikujejo glede na osebnostne značilnosti, spol, starost ali katerokoli drugo očitno značilnost.
2. Merjenje in statistično ocenjevanje sprememb posameznikov (a) v teku časa, (b) kot posledic različne vrste in obsega tretmaja, in (c) kot posledic istih tretmajev, ne glede na individualen nivo sposobnosti.

Merjenje in statistična ocena sprememb v skupinah

Koncept je najlažje ilustrirati s pomočjo diskusije o situaciji, v kateri želimo dokumentirati diferencialen vpliv eksperimentalnega tretmaja na visoko in nizko sposobne sodelujoče (kakor v našem primeru povprečnih managerjev glede na "superzvezdnike"). Uporaba metode pa, kot bomo videli, nikakor ni omejena zgolj na takšne primere. Kadar isti test - tudi če ni enodimenzionalen - služi oceni dosežka pred in po intervenciji, lahko vsako postavko, ki je bila podana na obeh testiranjih, obravnavamo kot par postavk z različnimi parametri, v sklopu običajne Raschevе lestvice (Raschev ali enoparametrski model), natančneje, kot "miniaturno" Raschevo lestvico z dolžino dve postavki. Npr. če podamo istih 10 postavk na pre- in post-testu, lahko tako oblikujemo 10 miniaturnih Raschevih lestvic. Ni potrebno, da te lestvice merijo skupno dimenzijo; lahko bi bile, kot tudi v primeru kliničnih študij tudi so, osnovane za merjenje 10 različnih dimenzij, z namenom kar se da celostnega ocenjevanja sprememb. Kljub temu ni potrebno uporabiti dolgih testov, ker vsaka postavka meri drugo latentno dimenzijo (te dimenzije lahko korelirajo ali so na kakšen drug način med seboj odvisne, ali neodvisne). Po tem lahko v drugem koraku ocenimo, ali je možno katere izmed učinkov generalizirati preko vseh postavk. Če je to možno (in iz mnogih dostopnih študij si zdi, da je temu pogosto tako), lahko podamo celostno oceno učinka tretmajev, ali drugače ocenimo relativno velikost učinka na različne vključene "dimenzije". Očitno je, da je rezultat sklop zelo fleksibilnih postopkov.

Enak postopek lahko uporabimo za identifikacijo posameznikov, ki so se spremenili. Splošen model predpostavlja enak učinek za vse sodelujoče za vsako izmed "miniaturnih Raschevih lestvic". Model je možno razširiti tako, da nam omogoča identifikacijo gruč ljudi, ki so podobni drugim v isti skupini (v smislu, da se močneje odzovejo na eksperimentalne spremenljivke), a se razlikujejo od tistih v drugih skupinah, ki se odzovejo drugače. Postopki so popolnoma vzporedni postopkom za identifikacijo gruč postavk, ki se odzivajo podobno kot druge znotraj njihove skupine in različno od postavk v drugih skupinah.

Kljub temu, da je razvoj teh postopkov osnovan na podlagi TOP, metoda v temeljih odstopa od predpostavke enodimenzionalnosti večine modelov TOP. Zaradi tega je predstavljeni model poimenovan kot "Linearni logistični model z manj zahtevnimi

predpostavkami - LLMMZP" (Linear Logistic Model with Relaxed Assumptions – LLRA; glej Fischer, 1995b). Vključen je v programski opremi LPCM-Win 1.0 (Fischer in Ponocny-Seliger, 1998). Na razpolago so tudi različice in razširitve metode za postavke z več kot dvema odgovornima kategorijama (bralec, ki ga zanima psihometrično ozadje tega pristopa, lahko več o njem najde v knjigi *Rasch Models*, katere urednika sta Fischer in Ponocny-Seliger (1998), in knjigi *Handbook of the Usage of LPCM Win 1.0*, Fischer in Ponocny-Seliger (1998), kratek formalni opis LLMMZP pa je možno najti tudi pod opombo 3).

LLMMZP in sorodni modeli za merjenje sprememb omogočajo več oblik raziskovalnega modela:

- Uporaba istih postavk v dveh ali več različnih točkah v času. Postavke so lahko enodimenzionalne, vendar to ni nujno.
- Uporaba različnih, morda prekrivajočih se vzorcev postavk iz enodimenzionalne skupine postavk (ki jih poznamo iz poprejšnje študije TOP) v dveh ali več točkah v času (taka oblika načrta dovoljuje, kakor že omenjeno v enem izmed primerov, primerjavo težjega re-testa z lažjim pre-testom). Možno je uporabiti eno ali več enodimenzionalnih skupin postavk, kar ponovno omogoča večdimenzionalnost celotne skupine postavk. V takem primeru je pomembno, da v vsaki točki v času uporabimo vsaj eno postavko iz vsake izmed enodimenzionalnih skupin postavk, kar zagotavlja merjenje vseh latentnih dimenzij v vsaki točki v času. Število vključenih latentnih dimenzij je omejeno zgolj s praktičnimi omejitvami dolžine testa.
- Postavke so lahko dihotomne (kakor pri večini testov sposobnosti) ali pa politomne (z ocenjevalnimi lestvicami, kakor pri mnogih osebnostnih ali kliničnih ocenjevalni lestvicah).
- Možna je uporaba neomejenega števila eksperimentalnih in kontrolnih skupin. Eksperimentalna skupina je, po definiciji, skupina posameznikov, ki v istem času odgovarja na isto podskupino postavk in je deležna iste intervencije oz. kombinacije intervencij.
- Podatki so lahko popolni ali nepopolni. V primeru manjkajočih podatkov (npr. manjkajoča opazovanja) je potrebna uvedba več eksperimentalnih skupin, saj morajo vsi sodelujoči znotraj ene eksperimentalne skupine odgovoriti na isto podskupino postavk.

Očitno naštete lastnosti eskperimentalnega načrta zaobsegajo širok obseg možnih študij. V primeru, da je študija smiselno oblikovana glede na aplicirane kombinacije tretmajev, omogoča omenjena uporaba metodologije TOP ocene parametrov učinka, kot tudi enega ali več učinkov trenda, ki predstavljajo vplive na spremembe, nepovezane z uvedenimi tremaji. Metoda omogoča tudi izračun testov pomembnosti in standardnih napak za parametre učinka. Nadalje omogoča LPCM-Win postavljanje in testiranje standardnih hipotez (posplošljivost učinkov tretmaja ali

stopnje spremembe preko tako podskupin postavk, kot tudi podskupin sodelujočih), kot tudi kopico prirejenih hipotez. Na tej točki, pred predstavitvijo uporabe konceptov, je smotrno razložiti konceptualni pomik, ki omogoča uporabo TOP za do sedaj težko rešljive probleme. Eden izmed temeljnih konceptualnih premikov je osnovan na uporabi parametrov postavk (ki oblikujejo miniaturne Rascheve lestvice, omenjene poprej v sestavku) kot indeksov spremembe znotraj posameznikov. Izraženo tehnično - ista postavka predstavlja, uporabljena v dveh različnih časih, par "virtualnih" postavk z različnimi parametri postavk. Razlika med parametri postavk para postavk postane indikator sprememb sodelujočih na latentni dimenziji postavk. Glede na predpostavke posplošljivosti sprememb preko latentnih dimenzij, izmerjenih s pomočjo različnih postavk, in preko posameznikov znotraj eksperimentalne skupine, prispeva vsak par virtualnih postavk k celotni informaciji o obsegu sprememb v skupini. S tem kombinacija vseh teh prispevkov omogoča merjenje in statistično ocenjevanje sprememb.

Ocena parametrov učinka z uporabo LLMMZP ne vključuje ocene parametrov postavk ali parametrov posameznikov. Podane so samo ocene parametrov spremembe (npr.: učinki tretmaja ali sprememb, ki so se zgodili v času). Izračun je v celoti osnovan le na tistih odgovornih kombinacijah, kjer je posameznik uspešno odgovoril na le eno izmed postavk v paru (t.i. miniaturna Rascheva lestvica). Odgovorne kombinacije z obema nepravilnima ali pravilnima odgovoroma ne prispevajo k informacij o spremembi in jih zato zanemarimo. Priporočljiva je maksimizacija rezultatov 1 (in ne 0 in 2) na vsaki izmed teh miniaturnih Raschevih lestvic (parov postavk). To lahko dosežemo z inteligentnim izborom postavk, ki tvorijo omenjene pare.

Merjenje in statistično ocenjevanje sprememb posameznikov (a) v času, (b) kot posledic različne vrste in obsega tratmaja, in (c) kot posledic istih tratmajev, ne glede na individualen nivo sposobnosti in (d) z različnimi vzorci sposobnosti in osebnostnih značilnosti.

K preučevanju individualnih sprememb vodi drugačna raziskovalna motivacija: klinični psihologi, na primer, se sprašujejo, ali je pacient po določenem obdobju tretmaja sposoben izboljšati svoj testni dosežek; pedagoški psiholog želi primerjati individualni napredek v določenem obdobju s povprečnim napredkom posameznikov iste starosti; aplikativne psihologe zanima sprememba osebnostne poteze po uvedbi določenega treninga ali programa osebnostnega razvoja… Morda pomembnejše vprašanje predstavlja vprašanje spremembe kot odziva na določen tretma - bodisi psihofarmakološki, bodisi izobraževalni - torej rezultat spremembe kot rezultat s konceptualno ali napovedno veljavnostjo (za opis v netehničnem jeziku glej Fischer in Prieler, 2000).

Prvi primer merjenja sprememb na spremembah posameznika v TOP je bil Embertsonov (1991) »Večdimentionalni Raschev model za učenje in spremembo - VRMUS« (Multidimensional Rasch model for Learning and Change – MRMLC). Vendar pa je bila v tem »novem modelu« le izračunana razlika v dveh parametrih osebe (podobno kot enostavna razlika med surovimi točkami), ki sta bila ocenjena po TOP v dveh točkah v času 1 in 2. Fischer je ta pristop kritiziral in opisal nov model za

spremembo posameznika (Fischer, 1995). Njegovi glavni argumenti proti VRMUS so bili, (a) da Embertsonova mera spremembe temelji na asimptotični distribuciji ocene parametra osebe, kar zahteva zelo dolge teste, pri Fischerjevem modelu pa je možno uporabljati teste običajne dolžine, (b) da Embertsonov model sicer da nekatere parametre učinka, ki pa jih da tudi Fischerjev tudi v pogojih, v katerih jih Embertsonov ne.

Pri Fischerjevem pristopu ni nujna ocena posameznikovih parametrov v dveh točkah v času; ocenjeni so sami parametri spremembe. Uporabljeni testi so včasih testi sposobnosti z dihotomnimi postavkami (kot je SPM Plus, glej Fischer in Prieler, 2000), včasih pa lestvice sestavljene iz več odgovornih kategorij, kot so »vedno«, "pogosto", "včasih", "nikoli". V vseh teh primerih je sedaj dostopna tudi metodologija TOP, ki omogoča merjenje individualnih sprememb in statistično oceno spremembe. Poudariti pa velja, da mora biti izbrana skupina postavk - za razliko skupinsko orientiranih raziskav - enodimenzionalna. To ni težko razložiti, saj: če se raziskava osredotoči na posameznike in če vsaka postavka meri drugo dimenzijo ter sta možna le dva diskretna odgovora na postavko, postane podajanje znanstvene ocene količine spremembe na vsakih izmed latentnih dimenziji nemogoče. Poleg omenjene omejitve enodimenzionalnosti izbrane skupine postavk je omenjena metodologija TOP do sedaj razvita zgolj za preučevanje sprememb v dveh točkah v času. Pri raziskavah, ki vključujejo več kot dve časovni točki, je potrebna ločena analiza parov točk.

Kljub temu obstaja velika fleksibilnost glede sestave testa uporabljenega v dveh različnih točkah v času: iz podane enodimenzionalne skupine postavk lahko za vsako točko v času uporabimo katerokoli podskupino postavk. Sodelujočemu lahko podamo iste postavke dvakrat ali pa uporabimo popolnoma drugačno podskupino postavk za pre- in post-test, podskupini postavk pa se lahko tudi delno prekrivata. Raziskovalec lahko v skladu s svojimi teorijami in raziskovalnimi cilji po želji oblikuje izbor postavk. Če raziskovalec, na primer, pričakuje porast rezultata sposobnosti ali poteze sodelujočega, lahko za pre-test izbere lažje postavke kot za post-test, s čimer pričakovani premik rezultata na latentni dimenziji kompenzira z večjo težavnostjo postavk.

Ideja za to psihometrično metodo je, da se količina spremembe pri preučevanem posamezniku kaže v parametrih postavk: koncept "virtualnih" postavk tako ponovno postane ključen za razumevanje pristopa. Namesto pojmovanja spremembe kot spremembe posameznikovih parametrov (sposobnosti), si je smotrno zamisliti spremembo kot premik post-test parametrov postavk glede na parametre postavk pre-testa. Torej je posameznikov parameter (sposobnosti) - kljub resnični spremembi - tehnično možno razumeti kot konstanten, medtem ko parametre postavk post-testa zamenjajo "virtualni" parametri postavk. Posledično lahko z odgovori sodelujočega na obeh testih ravnamo kot z odgovori sodelujočega na enem testu, katerega dolžina je enaka vsoti dolžin pre- in post-testa. To nam omogoči uporabo tako imenovane "metode pogojnega največjega verjetja - PNV". (conditional maximun likelihood method). Prednost metode predstavlja izključitev posameznikovih parametrov iz nadaljnjih korakov ocenjevanja in postopkov statističnega testiranja. Metoda se izogne

tudi kakršnimkoli asimptotičnim približkom, saj sloni na uporabi točne verjetnostne distribucije rezultatov prirastka.

Tu ni mesta za podroben formalen opis metodologije. Zainteresiranega bralca naj napotimo na delo Fischerja (1995a, 2000), podali pa smo tudi krajši oris pod končno opombo 4. Omeniti velja, da so podobne metode v zvezi s preučevanjem "prileganju posameznika" v Raschevem modelu predlagali tudi Klauer (1991), Liou (1993) in Ponocny (2000). Dovolj je, da povemo, da metoda za vsakega sodelujočega nudi oceno spremembe na latentni dimenziji, da je ta mera neodvisna od prave začetne stopnje sposobnosti ali poteze, da je možno izračunati stopnje zaupanja za pravo individualno količino spremembe, in da je možno testirati statistično pomembnost količine spremembe.

Zanimivo razširitev LLMMZP / LLTM modelov za ocenjevanje sprememb, ki je omenjena tu, predstavi Meiser (1995). Alternative LPCM-Win programski opremi, ki je osnovana na pristopu pogojnega največjega verjetja – PNV, so programski paketi osnovani na pristopu mejnega največjega verjetja - MNV (marginal maximum likelihood) Wuja, Adamsa in Wilsona (1995) in programski paket, osnovan na linearnem modelu za količnike logaritmske verjetnosti (Linacre, 1996). Primere uporabe te metodologije v namene preseganja teh težko zasledljivih težav lahko najdemo v Apfelthaler (2000), Erasim (1995), Fischer in Seliger (1997), Jenull-Schiefer (2000), Spiel in Glück (1998), Stögerer (2000) in Wernsdorf (1998). Na tem mestu je dovolj, da povemo nekaj več o že omenjeni študiji avstrijskih vojakov Prielerja (2000) o napovedni veljavnosti točk, ki merijo odgovor na stres. Uporabljena je bila obsežna baterija psiholoških testov pred in po napornem nočnem pohodu z namenom odkrivanja katere spremembe rezultatov bodo najbolje napovedovale nadaljnji uspeh kandidatov kot oficirjev. Izsledki so bili sledeči:

- Napovedna veljavnost tako pre- kot re-test rezultatov je bila v primerjavi z napovedno veljavnostjo rezultatov sprememb (prirastek ali upad) nizka, ne glede na to, ali so ti rezultati bili rezultati visoko, srednje ali nizko sposobnih.
- Trije od sedmih rezultatov sprememb (ali razlik) so omogočili prepoznavo tistih, ki ne bi dokončali enoletnega programa oficirskega usposabljanja, kar je omogočilo nove, visoko veljavne kriterije za izbor kadrov.
- Napovedna veljavnost nekaterih izmed rezultatov sprememb je bila tako visoka, da je bilo mogoče zgolj na podlagi teh rezultatov iz programa izločiti vse kandidate z visoko negativnimi rezultati spremembe.
- Določene postavke niso imele napovedne veljavnosti. Zato je bilo testno baterijo mogoče skrajšati.

Avstrijska vojska je zdaj prva institucija na svetu, ki uporablja to metodo kot standarden postopek ocenjevanja posameznikovih sposobnosti. Na to lahko gledamo kot na nov kvalitativen preskok k dobro znanemu kriteriju v selekciji kadrov »pravega človeka na pravo mesto«. Testiranje pod pritiskom v povezavi z novim pristopom

merjenja zagotavlja dober kadrovski management, kar v veliki meri zmanjša stroške kadrov, in v veliki meri zveča motivacijo vsakega posameznega vojaka. V svoji raziskavi v nemški vojski je Melter (1992) ugotovil, da »je že v postopku selekcije nekaj malega informacije o kasnejši prekinitvi vojaške kariere«. Z uporabo postopkov, ki so bili opisani, je zdaj že v začetni fazi novačenja lažje izslediti vojake, ki bi lahko kasneje prekinili s kariero. Z drugimi besedami – zdi se, da je zdaj iz obstoječih podatkov mogoče izvedeti več kot je bilo poprej.

Zaključek

Opisali smo vrsto vse večjih težav, ki otežujejo psihološko raziskovanje in – v določenih primerih – pripeljejo do tega, da raziskovalci prihajajo do zelo zavajajočih zaključkov, ki so imeli v preteklosti že velike učinke na stroko. Nekatere izmed teh težav so dobro poznane že od začetka uporabe statistike, čeprav jih (kot je pozno opozorila delovna skupina APA za statistično pomembnost) raziskovalne skupine v svoji zagretosti za uporabo sofisticiranih metod - preden se dodobra spoznajo s svojimi podatki - pogosto prezrejo. Drugo skupino težav, ki smo jih opisali, je prepoznala relativno majhna skupina metodologov v zadnjih štiridesetih letih. Kljub temu, da so bile prepoznane, pa zanje še ni bilo uporabnih rešitev. Posledica tega je bila, da so bili najbolj teoretično usmerjeni praktiki, kadar so bili negotovi v svojem bežnem poznavanju problematičnih lastnosti svojih metod, prepričani, da morajo nekaj narediti in so nadaljevali z uporabo enostavnih postopkov, ne da bi se zavedali, kako resno so vključene napake merjenja spodkopale možnost, da bi lahko prišli do kakršnihkoli smiselnih zaključkov. Tem teoretično orientiranim raziskovalcem se je pridružila množica raziskovalcev, ki so bili pred praktično nalogo demonstracije diferencialne vrednosti različnih odmerkov različnih drog, primerjanja učinka različnih terapij za »različne vrste ljudi« znotraj različnih zdravstvenih programov, ali določanja, kateri izmed številnih terapevtskih načinov (z različno ceno) prinaša največji napredek pri posameznem pacientu. K tem smo dodali še druge resne težave, ki izhajajo iz dejstva, da večina evalvacij, ki so bile izvedene do sedaj (bodisi na posamezniku ali skupinah), ni dovolj celostnih in iz njih zato sledijo zavajajoči zaključki in napačni ukrepi. Zanimivo je, da, čeprav se teh težav verjetno zaveda, delovna skupina APA ni opozorila nanje.

Po tem, ko smo opisali nekatere izmed težav, ki se pojavljajo pri merjenju sprememb pri skupinah in posameznikih, smo povzeli novo, genialno, teoretično osnovano in praktično rešitev ali – bolje – niz rešitev teh težav. Čeprav te rešitve niso nove v drugih pogledih, gre za novost z vidika računalniških programov, ki so potrebni za njihovo izvedbo in so postali dostopni šele pred kratkim.

Dve, morda po naključju odkriti prednosti teh postopkov smo hranili za konec. Omogočajo nam, da ugotovimo, ali se je določen posameznik pomembno odzval na določeno zdravilo, terapijo, izobraževalni ali razvojni program (in če, na kakšen način). S tem, ko lahko skrajšamo dolžino testov, za katero se včasih zdi, da je nujna, nam

olajšajo osredotočanje na celostnost v evalvaciji ljudi in programov,

Na koncu je potrebno dodati opozorilo. Opisani postopki niso zdravilo vseh ran. Sočasno ocenjevanje parametrov postavke in učinkov tretmaja (kar je možno z uporabo intraindividualnega modela hibridne LLMMZP za razlike med skupinami) je lahko velik vir pristranosti. Zato je, kadar je to mogoče, treba uporabiti enega izmed dveh pristopov: ali je potrebno določiti parametre postavk v ločeni kalibracijski študiji (na primer uporaba testov TOP, za katere so dosegljivi parametri težavnosti postavk) ali pa uporabljati isto postavko ponovno. Ne smemo pa pozabiti (čeprav bi želeli, da nam tega ne bi bilo treba napisati), da je dobro zasnovan načrt postopka nujen, če želimo dobiti smiselne rezultate.

Priloga 1 (opombe)

1. Natančneje – razlike med surovimi dosežki, ki so v povezavi z razlikami v latentni lastnosti, močno variirajo: (a) s splošno stopnjo težavnosti testa v primerjavi s merjeno lastnostjo in (b) z distribucijo parametrov postavke glede na interval latentne poteze, na katerem se zgodi sprememba (Fischer in Prieler, 2000), oziroma z delom in obliko tega, čemur Fischer (1991) pravi »karakteristična krivulja testa«, s katerim je merjena razlika. To drži tudi v primeru, testov, ki zadostijo kriterijem najpopularnejših različic modelov TOP – torej drži pri testih, ki se skladajo z enoparametrskim (model, ki ne dovoli razlike v naklonu ali ugibanju), dvoparametrskim (model, ki dopušča razlike v naklonu, ne pa v ugibanju) in triparametrskim (model, ki dopušča razlike tako v naklonu, kot tudi v ugibanju) modelom TOP (glej Hambleton, Swaminathan in Rogers, 1991; van der Linden in Hambleton, 1997 za poglobljen opis teh modelov). Bolj natančno – enodimenzionalnost, ki jo zagotavlja skladanje z večino modelov TOP (Hambleton, Swaminathan in Rogers, 1991; van der Linden in Hambleton, 1997) sama po sebi ne zagotavlja neizkrivljenosti točk spremembe.

2. Naj povzamemo Fischerja in Molenaarja (1995): V psihološkem testu ali testu stališč skušamo meriti stopnjo lastnosti, ki jo ima posameznik. Družbene znanosti in znanosti o vedenju imajo pogosto opravka z lastnostmi kot so inteligentnost, računske spretnosti, nevroticizem, politični konzervativizem ali spretnost prstov. Lahko je najti primere opazljivega vedenja, ki kaže na to, da ima posameznik več ali manj take splošne lastnosti, vendar pa ima koncept dodano vrednost v tem smislu, da ga nobeno izmed manifestnih vedenj ne zajema v celoti. Zato takšnim lastnostim pravimo latentne poteze. Uporaba testa ali lestvice predpostavlja, da lahko neposredno sklepamo na posameznikovo izraženost latentne poteze na osnovi njegovih/njenih odgovorov na niz dobro izbranih postavk, kar nam tudi omogoča, da napovedujemo njegovo/njeno vedenje, ko bo soočen s postavkami iz istega področja. Statistični model merskega procesa bi nam moral omogočati takšne napovedi. Še več, omogočati nam bi moral posploševanje na druge osebe, ki bi reševale isti test, in odgovoriti na vprašanja o posplošljivosti na različne situacije testiranja.

Opaženo vedenje so testni dosežki, v klasični teoriji sta neodvisni spremenljivki latentna spremenljivka in napaka. Odvisna spremenljivka je opažen dosežek na testu:

Opažen dosežek = Pravi dosežek + Napaka

Vendar pa ima ta »Teorija pravega dosežka« veliko pomanjkljivosti za konstrukcijo testov (Hambleton, Swaminathan in Rogers, 1991). Kot prvo – težavnost postavke in njena diskriminativnost sta odvisni od skupine. Težava se pojavi, ko vzorec testirancev ne odraža populacije, ki ji je test namenjen, kar vodi do tega, da bodo parametri, dobljeni na njem, omejeni v uporabnosti. Kot drugo – ocene lastnosti posameznikov, ki so odvisne od postavk, se nanašajo na določeno izbiro postavk v testu. To oteži primerjavo, kadar posamezniki rešujejo različne teste. Kot tretje – standardna napaka merjenja naj bi bila enaka za vse testirance, kar pa ni verjetno, saj dosežki na posameznem testu niso enako natančne mere testirancev z različno izraženo lastnostjo. Kot četrto – zanesljivost je opredeljena kot korelacija med testnimi dosežki na »vzporednih testih«, ki pa jih je v praksi težko izdelati.

Zaradi tega so psihometriki iskali alternativne teorije in modele merjenja. TOP je presegla omejitve klasične testne teorije. V TOP je opaženo vedenje posamezen odgovor na postavko (na primer 0 ali 1). Latentna spremenljivka vpliva na verjetnost odgovora na te postavke. Verjetnost, da bo posameznik uspešno rešil ali se strinjal s posamezno postavko, je odvisna od njegove stopnje izraženosti poteze in težavnosti postavke na naslednji način:

Verjetnost (pravilne rešitve) = Funkcija [(stopnja poteze – težavnost postavke)]

Raschev model TOP je sledeča enostavna logistična funkcija stopnje poteze in težavnosti postavke:

$$P(X_{ij}=1/q_j, b_i) = \exp(q_{vi}-b_i)/[1+\exp(q_{vi}-b_i)]$$

Pri čemer je q_j posameznikova izraženost poteze, b_i pa stopnja teževnosti postavke (Rasch, 1960; Embertson, 1999).

Čeprav bi večina avtorjev v principu želela omejiti uporabo fraze »prileganje Raschevemu modelu« na teste, ki se skladajo z gornjim enoparametrskim logističnim modelom, kažejo vroče razprave o tem, ali se Ravnove zahtevne matrice skladajo z »Raschevim modelom«, da v praksi ni vedno tako. Nizi karakterističnih krivulj postavk RPM, ki so jih izrisali in objavili v veliko študijah, izvedenih v različnih državah od 1942 (glej Raven, 1981), jasno kažejo na to, da se krivulje razlikujejo v naklonih, in da parameter ugibanja vpliva prej kot začnejo postavke diskriminirati. Iz tega sledi, da je ustrezen le triparametrski model in puristična interpretacija zahtev za skladanje z »Raschevim modelom« ne more biti zadovoljena. Ne samo, da je Rasch testiral svoj model na ta način, da je pokazal, da se prilega Ravnovim zahtevnim matricam, tudi drugi raziskovalci (kot na primer Andrich), ki so izvedenci tako v TOP kot tudi v Raschevih modelih, so trdili isto. Iz tega sledi, da puristična interpretacija tega, kar naj bi bil »Raschev model«, v praksi ni bila vedno osvojena. Hambelton je z muko (pri čemer mu strokovna javnost ni vedno prisluhnila) izpostavljal, da gre pri teh razpravah za dve vprašanji: (a) narava in stopnja *kriterija*, ki je bil postavljen za sprejem nečesa kot skladnega z »Raschevim modelom«, in (b) narava in velikost obravnavanih vzorcev… saj se izkaže, da so parametri testa in postavke, ki jih dobimo na osnovi majhnih vzorcev, zelo nestabilne, še posebej pri triparametrskih modelih. V tem članku uporabljamo izraz »Raschev model« za vsak test, ki se sklada z eno-, dvo– ali

triparametrskim modelom.

3. Pričujoči kratek formalen opis LLMMPZ je povzet iz Fischer in Ponocny-Seliger (1998). Denimo, da imamo test s k postavkami I_i, ki ga reši niz oseb S_v na dveh točkah v času T_1 in T_2 – pred in po določenem tretmaju. Naj bodo verjetnosti pozitivnega odgovora na postavko I_i:

$$P(+/S_v, I_i, T_1) = \exp(q_{vi})/[1 + \exp(q_{vi})]$$

v času T_1 in

$$P(+/S_v, I_i, T_1) = \exp(q'_{vi})/[1 + \exp(q'_{vi})]$$

v času T_2.

Parameter q_{vi} izraža položaj posameznika S_v v času T_1 na latentni dimenziji D_i, ki jo merimo s postavko I_i. q_{vi} lahko prosto variirajo, tako da lahko postavke merijo neodvisne, korelirane ali kako drugače soodvisne poteze. Tako nimamo omejitev glede povezav med potezami, kar nam omogoča, da lahko model uporabljamo na zelo različnih področjih. Podobno parameter q'_{vi} izraža položaj posameznika S_v na dimenziji D_i v času T_2.

Da bi resnično dobili veljavne rezultate, mora biti, če je test normalne dolžine (ima več kot 10 postavk), v vsaki izmed skupin udeležencev posameznega tretmaja vsaj 30 posameznikov. Pravzaprav je veljavnost testnih rezultatov odvisna od produkta n*k. Iz tega sledi, da je teoretično možno imeti manj kot 30 ljudi, vendar pa mora imeti v tem primeru test nerazumljivo veliko število postavk. Pomanjkljivost te metode je v tem, da dobimo neustrezen rezultat, če ne obstaja posplošljivost med deli postavk ali osebami na sploh (to pomeni, da je učinek tretmaja različen za vsako posamezno postavko ali osebo; rezultat, ki običajno ni interpretabilen).

4. Sledeč kratek formalen opis LLTM za opazovanje sprememb posameznikov je povzet iz Fischer (2001). Naj bo vzorec postavk sestavljen iz postavk $I = I_1,....,I_k$. Domnevamo, da se Model delnih dobičkov - MDD (Partial Credit Model - PCM) prilega tem postavkam:

$$P(X_{vij} = 1/ q_{vi}, b_{i0},....., b_{imi}) = \exp(jq_v + b_{ij}) / S_{l=0} \exp(l/q_v + b_{il})$$

Pri čemer:
X_{vij} predstavlja naključno odgovorno spremenljivko z vrednostjo $x_{vij} = 1$, če je posameznik S_v na postavko I_i pripada kategoriji odgovorov C_{ij}, sicer pa je $x_{vij} = 0$;
pri tem je m_i število kategorij postavk minus 1 (kategorije odgovorov so oštevilčene z $0,\ldots,m_i$);
q_v je parameter posameznika S_v;
b_{ij} za $j = 0,\ldots m_i$, so parametri kategorije postavke I_i.

Da bi bil model identifikabilen, moramo zagotoviti $k + 1$ normalizacij na parametrih. Bodite pozorni na to, da sta Model lestvice stopnje – MLS (Rate Scale Model - RSM) in Raschev model posebna oblika MDD.

Gornji model naj bi veljal za vse postavke, če jih uporabimo le enkrat.

Tako model velja za postavke pre-testa I_1. Za post-test I_2 pa moramo model spremeniti,

saj domnevamo, da je potrebno parameter posameznikaq_v zamenjati z $q_v + h_v$, pri čemer je h_v parameter, ki predstavlja stopnjo spremembe pri posamezniku S_v. Za oceno parametra spremembe h_v uporabimo metodo pogojne največje verjetja – PNV (conditional maximum likelihood method - CML).

Če torej obstaja test, ki se prilega PCM (ali RSM ali Raschevemu eno-, dvo- ali triparametrskemu modelu), lahko izračunamo tabelo za pomembne spremembe na nivoju posameznika med dvema točkama v času.

Pomanjkljivost metode je v tem, da pri osebah z vsemi ali nobeno doseženo točko v dveh časovnih točkah ne moremo zaključevati ničesar o resnični spremembi latentne poteze.

Literatura

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 561–573.

APA Task force on Statistical Inference (1999). Glej: Wilkinson, L. in Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594-604.

Apfelthaler, E. (2000). Medikamentöse und psychotherapeutische Effekte in der Behandlung der erektilen Dysfunktion. (neobjavljena disertacija). Dunaj: University of Vienna.

Bereiter, C. (1963). Some persisting dilemmas in the measurement of change. V C.W. Harris (ur.), *Problems in measuring change* (str. 3-20). Milwaukee, WI: University of Wisconsin Press.

Bouvier, U. (1969). *Evolution des Cotes a Quelques Tests.* Belgium: Centre de Recherches, Forces Armees Belges.

Budoff, M. (1973). Measuring learning potentials: An alternative to the traditional intelligence test. *Studies in Learning Potentials, 3,* 39.

Budoff, M., Corman, L. in Gimon, A. (1976). An educational test of learning potential assessment with Spanish speaking youth. *Inter-American Journal of Psychology, 10,* 13–24.

Carver, R.P. (1989). Measuring intellectual growth and decline. *Psychological Assessment, 1(3),* 175–180.

Cronbach, L.J. in Furby, L. (1970). How should we measure "change" – or should we? *Psychological Bulletin, 74*, 68-80.

Embretson, S.E (1991). Implications of a multidimensional latent trait model for measuring change. V L.M. Collins in J.L. Horn (ur.), *Best Methods for the Analysis of Change.* American Psychological Association.

Erasim, U. (1995). Anwendung des LLRA: "Der Einfluss von mentalem Training auf die sportliche Leistung jugendlicher Tennisspieler". (neobjavljena disertacija). Dunaj: University of Vienna.

Feuerstein, R. (1979). *The Dynamic Assessment of Retarded Performers*. Baltimore: University Park Press.

Feuerstein, R., Klein, P. in Tannenbaum, A. (ur.). (1990). *Mediated learning experience: Theoretical, psycho-social, and educational implications*. Proceedings of the First International Conference on Mediated Learning Experience. Tel Aviv: Freund.

Fischer, G.H. (1972). A measurement model for the effect of mass-media. *Acta Psychologica, 36,* 207-220.

Fischer, G.H. (1974). *Einführung in die Theorie psychologischer Tests*. [Introduction to mental test theory.] Bern: Huber.

Fischer, G.H. (1977). Linear logistic latent trait models: Theory and application. V H. Spada in W.F. Kempf (ur.). *Structural Models of Thinking and Learning.* Bern: Huber.

Fischer, G.H. (1983). Logistic latent trait models with linear constraints. *Psychometrika, 46*, 59-77

Fischer, G.H. (1989). An IRT-based model for dichotomous longitudinal data. *Psychometrika, 54*, 599-624.

Fischer, G.H. (1991). A new methodology for the assessment of treatment effects. *Evaluación Psicológica/Psychological Assessment, 7(2),* 117–147.

Fischer, G.H. (1995). Some neglected problems in IRT. *Psychometrika, 60,* 459–487.

Fischer, G.H. (1995). Linear logistic models for change. V G.H. Fischer in I.W. Molenaar (ur.), *Rasch models, recent developments and applications* (str. 158-180). New York: Springer.

Fischer, G.H. (1997). Multidimensional linear logistic models for change. V W. J. van der Linden in R. K. Hambleton (ur.), *Handbook of modern Item Response Theory* (str. 323–346). New York: Springer.

Fischer, G.H. (2001). Gain scores revisited under an IRT perspective. V A. Boomsma, M.A.J. van Dujin in T.A.B. Snijders (ur.), *Essays on Item Response Theory* (str. 43-68). New York: Springer.

Fischer, G.H. in Molenaar, I.W. (ur.). (1995). *Rasch models, recent developments and applications.* New York: Springer.

Fischer, G.H. in Seliger, E. (1997). Multidimensional linear logistic models for change. V W.J. van der Linden in R.K. Hambleton (ur.), *Handbook of modern Item Response Theory* (str. 323-346). New York: Springer.

Fischer, G.H. in Ponocny-Seliger, E. (1998). *Structural Rasch modeling.* Handbook of the usage of LPCM-Win 1.0, ProGAMMA.

Fischer, G.H. in Prieler, J.A. (2000). *An IRT-based methodology for the assessment of change.* Appendix 2 v J. Raven, J.C. Raven in J.H. Court, *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 3: The Standard Progressive Matrices*. Oxford: Oxford Psychologists Press.

Flynn, J.R. (1984). The mean IQ of Americans: massive gains 1932 to 1978. *Psychological Bulletin, 95,* 29–51.

Flynn, J.R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101,* 171–191.

Flynn, J.R. (1999). Searching for justice: The discovery of IQ gains over time. *American Psychologist, 54(1),* 5–20.

Flynn, J.R. (2000). IQ gains, WISC subtests and fluid ***g***: ***g*** theory and the relevance of Spearman's hypothesis to race. V G.R. Bock, J.A. Goode in K. Webb (ur.), *The nature of intelligence* (str. 202-227). Chicheser: Wiley.

Guthke, J. (1982). The learning test concept – An alternative to the traditional static intelligence test. *German Journal of Psychology, 6,* 306–324.
Guthke, J. in K.H. Wiedl (1996). *Dynamisches Testen.* Gottingen: Hogrefe.
Hambleton, R.K. (1988). Comments made in a symposium on *Questionable Assumptions in Test Construction*, held at the meeting of the International Association for Applied Psychology, Sydney.
Hambleton, R.K. (1989). Constructing tests with item response models: A discussion of methods and two problems. *Bulletin of the International Test Commission, No.28/29*, 96–106. Strasbourg: ITC.
Hambleton, R.K., Swaminathan in Rogers (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage.
Harris, C. W. (1963). *Problems in measuring change.* Madison, Milwaukee, WI: University of Wisconsin Press.
Jenull-Schiefer, E. (2000). Evaluation verhaltenstherapeutischer Gruppentherapieprogramme zur Verbesserung sozialer Fertigkeiten bei schizophren Erkrankten (neobjavljena disertacija). Dunaj: University of Vienna.
Holtzmann, W.H. (1963). Statistical models for the study of change in the single case. V C.W. Harris (ur.), *Problems in Measuring Change* (str. 199-211). Milwaukee, WI: University of Wisconsin Press.
Klauer, K.C. (1991). An exact and optimal standardized person test for assessing consistency with the Rasch model. *Psychometrika, 56,* 213–228.
Linacre, J.M. (1996). Generalizability theory and many-facet Rasch measurement. V G. Engelhard, jr. in M. Wilson (ur.), Objective measurement. Theory into practice, Vol. 3 (str. 85-98). Norwood, NJ: Ablex.
Liou, M. (1993). Exact person tests for assessing model-data fit in the Rasch model. *Applied Psychological Measurement, 17,* 187–195.
Lord, F.M. (1963). Elementary models for measuring change. V C.W. Harris (ur.), *Problems in measuring change.* Madison, WI: University of Wisconsin Press.
Lord, F.M. in Novick, M.R. (ur.) (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.
Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149-174.
Meiser, T. (1996). Loglinear Rasch models for the analysis of stability and change. *Psychometrika, 61,* 629-645.
Melter, A.H. (1992). Psychologische Untersuchung der Offiziersanwärter mit nicht erfolgreicher Offiziersaus-bildung. V Rauch, M. (ur.), *Jahrbuch des psychologischen Dienstes der Bundeswehr,* München: Verlag für Wehrwissenschaften.
Molenaar, I.W. (1995). Some background for Item Response Theory and the Rasch model. V G.H. Fischer in I.W. Molenaar (ur.), *Rasch models, recent developments and applications* (str. 3-14). New York: Springer.
Ponocny, I. (2000). Exact person fit indexes for the Rasch model for arbitrary alternatives. *Psychometrika, 65,* 75–106.
Prieler, J. A. (1998). *Validation of personnel selection in the Austrian army*. Paper presented at the International Congress of Applied Psychology, San Francisco.
Prieler, J. A. (2000). *Evaluation eines Ausleseverfahrens für Unteroffiziere beim Österreichischen Bundesheer* (Validation of personnel selection of officers in the Austrian Army) (neobjavljena disertacija). Dunaj: University of Vienna.

Rasch, G. (1947). Quoted by B.D. Wright in a foreword to Rasch, G. (1980). *Probabalistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* Chicago, IL: University of Chicago Press.

Raven, J. (1981). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Research Supplement No.1: The 1979 British Standardisation of the Standard Progressive Matrices and Mill Hill Vocabulary Scales, Together With Comparative Data From Earlier Studies in the UK, US, Canada, Germany and Ireland.* Oxford: Oxford Psychologists Press.

Raven, J. (1989). Questionable assumptions in test construction. *Bulletin of the International Test Commission, 28-29,* 67–95.

Raven, J. (1991). *The tragic illusion: Educational testing*. Unionville, NY: Trillium Press.

Raven, J. (1994). *Managing education for effective schooling: The most important problem is to come to terms with values*. Unionville, NY: Trillium Press.

Raven, J. (1995). *The new wealth of nations: A new enquiry into the nature and origins of the wealth of nations and the societal learning arrangements needed for a sustainable society*. Unionville, NY: Royal Fireworks Press.

Raven, J. (2000a). Ethical dilemmas. *The Psychologist, 13,* 404–406.

Raven, J. (2000b). The Raven's *Progressive Matrices*: Change and stability over culture and time. *Cognitive Psychology, 41,* 1–48.

Raven, J., Raven, J.C. in Court, J.H. (1998a). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 1: General Overview.* Oxford, England: Oxford Psychologists Press.

Raven, J., Raven, J.C. in Court, J.H. (1998b). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 4: The Advanced Progressive Matrices.* Oxford, England: Oxford Psychologists Press.

Raven, J., Raven, J.C. in Court, J.H. (2000). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 3: The Standard Progressive Matrices.* Oxford, England: Oxford Psychologists Press.

Raven, J. in Stephenson, J. (ur.). (2001). *Competence in the learning society.* New York: Peter Lang.

Raven, J.C. (1965). *Advanced Progressive Matrices, Sets I and II: Plan and use of the scale with a report of experimental work carried out by G.A. Foulds in A.R. Forbes.* London: H.K. Lewis.

Rost, J. (1996). *Lehrbuch Testtheorie, Testkonstruktion.* Bern: Huber.

Schaie, K.W. in Willis, S.L. (1986). *Adult Development and Ageing* (2nd edition). Boston: Little Brown.

Spearman, C. (1926). *Some Issues in the theory of g (including the law of diminishing returns)*. Address to the British Association Section J – Psychology, Southampton, England, 1925. Bound in Collected Papers, Psychological Laboratory, University College of London.

Spiel, C. in Glück, J. (1998). Item response models for assessing change in dichotomous items. *International Journal of Behavioral Development, 22(3),* 517–536.

Stögerer, P. (2000). Prognostische Bedeutung von "Angst vor Kontrollverlust" für die Therapie des Paniksyndroms (neobjavljena disertacija). Dunaj: University of Vienna.
Teasdale, T.W. in Owen, D.R. (1989). Continuing secular increases in intelligence and a stable prevalence of high intelligence levels. *Intelligence, 13,* 255–262.
Thorndike, R.L. (1975). *Mr. Binet's Test 70 Years Later.* Presidential Address to the American Educational Research Association.
Van der Linden, W.J. in Hambleton, R.K. (1997). *Handbook of modern Item Response Theory*. New York: Springer.
Wernsdorf, T. (1998). Konzentrative Bewegungstherapie und Ich-Erleben. (neobjavljena disertacija). Dunaj: University of Vienna.
Williams, R.H. in Zimmermann, D.W. (1996). Are simple gain scores obsolete? *Applied Psychological Measurement, 20,* 55–69.
Wright (1968). Sample free test calibration and person measurement. *Proceedings of the 1967 invitational conference on testing problems.* Princeton, NJ: Educational Testing Service.
Wright, B.D. in Masters, G.N. (1982). *Rating Scale Analysis. Rasch Measurement.* Chicago: Mesa Press.
Wu, M.L., Adams, R.J. in Wilson, M.R. (1995). *MATS: Multi-Aspect Test Software computer program.* Melbourne: Australian Council for Educational Research.

Prispelo/Received: 10.01.2002
Sprejeto/Accepted: 10.01.2002