Виолета ПЕТРОСКА — БЕШКА

ДИФЕРЕНЦИЈАЛНО ПОНДЕРИРАЊЕ КАЈ ТЕСТОВИТЕ ОД ТИПОТ ПОВЕЌЕЧЛЕН ИЗБОР

Тестовите од типот повеќечлен избор се состојат од ајтеми (задачи) кои покрај даденото прашање содржат и две или повеќе понудени можни одговори (алтернативи) од кои најчесто само еден е точен. Се нарекуваат тестови од типот повеќечлен избор затоа што присуството на варијаблата која се мери со тестот (обично е тоа знаење од некоја област или определена способност) го регистрираат преку избирање на една од повеќето понудени алтернативи. Изборот на точната алтернатива укажува на поседување на варијаблата во определен степен; изборот на една од неточните алтернативи (таканаречени дистрактори) е знак на отсуство на определен индикатор на таа варијабла.

Конвенционалниот начин на бодување на одговорите од типот повеќечлен избор се врши главно на два начина. Според едниот, најзастапениот начин, изборот на точниот одговор на ајтемот се наградува со 1 бод, а изборот на едниот од дистракторите се третира исто како и неизвршен избор и се бодува со 0 бодови. Тест-скорот (резултатот на тестот во целина) се определува едноставно како збир од јатемите на кои е даден точен одговор.

Според другиот начин на конвенционално бодување, ајтем-скорот (бодовите доделени на одговорот на ајтемот) може да изнесува 1 ако е избрана точната алтернатива, — $\frac{1}{к-1}$ ако е избран еден од дистракторите (при што „к" го означува бројот на понудените алтернативи на ајтемот) или 0 ако не е извршен никаков избор. Според овој начин, тест-скорот (S) се определува со примена на таканареченото бодување по формула (formula scoring) и претставува разлика помеѓу бројот на точно одговорени ајтеми (R) и бројот на неточно одговорени ајтеми (W) намален к-1 пати:

$$S = R - \frac{W}{к-1}$$

На неточно избраните одговори им се доделуваат негативни бодови со оправдување дека на тој начин се врши корекција за погодување на точниот одговор кое е можно кога се користат ајтеми од типот повеќечлен избор.

Она што е заедничко за двата конвенционални начини на бодување на тестовите од типот повеќечлен избор е што точниот одговор на сите ајтеми се наградува подеднакво, со по 1 бод. Овој принцип на бодување се засновува на претпоставката дека сите ајтеми во тестот имаат подеднаква вредност како индикатори на варијаблата која се мери. Се поставува прашањето колку е таквата претпоставка одржлива, односно, дали навистина сите точно одговорени ајтеми укажуваат на еднаков степен на поседување на мерената варијабла.

Истовремено, доделувајќи 0 бодови (според едниот) или $\frac{1}{к-1}$ бодови (според другиот начин) на ајтемите кај кои е избрана една од погрешните понудени алтернативи, конвенционалните начини на бодување подеднакво ги третираат сите дистрактори во рамките на секој од ајтемите. Примената на ваквиот принцип на бодување потекнува од претпоставката дека сите дистрактори имаат еднаква вредност, при што не се зема предвид фактот дека некои дистрактори се поблиски, а други се подалечни од точниот одговор. Ваквата претпоставка е дискутабилна — се поставува прашањето колку е оправдано да се смета дека поблиските (помалку неточните) дистрактори укажуваат на ист степен на отсуство на конкретниот индикатор на мерената варијабла како и подалечните (повеќе неточните) дистрактори.

Потребата од надминување на ваквите недостатоци во изразувањето на ајтем-скорот по конвенционалните начини на бодување на одговорите кај тестовите од типот повеќечлен избор резултирала во развивање на техниките на диференцијално пондерирање. Постапките на диференцијално пондерирање се разликуваат од конвенционалните начини на бодување по тоа што припишуваат различни пондери (различни нумерички вредности, односно различен број бодови) на ајтемите или на алтернативите на секој ајтем.

Кога се применува диференцијално пондирање на ајтемите, секој ајтем се вредунува со одреден број бодови кои се определуваат во зависност од тоа колкав е придонесот на ајтемот во утврдување на присуството на варијаблата која се мери со тестот. Кога се применува диференцијално пондирање на алтернативите, за секоја алтернатива во рамките на секој ајтем се утврдува определена бодовна вредност зависно од тоа дали алтернативата е точниот одговор, дали е поблизок или е подалечен дистрактор.

Со постапките на диференцијално пондерирање тест-скоровите се определуваат врз основа на пондерите кои им се припишуваат на ајтемите или на ајтем-алтернативите пред почетокот на самото бодување. Пондерите, пак, можат да бидат определени на два различни начина: логички (априористички) и емпириски. Кај логичкото пондерирање износот на бодовите за секој од ајтемите или за секоја од алтернативите се определува однапред. врз основа на проценетата вредност на ајтемите, односно алтернативите. Кај емпириското пондерирање износот на бодовите се утврдува врз база на тоа колкав е уделот на ајтемот, односно алтернативата во општите психометриски карактеристики на тестот за определен примерок испитаници.

Од постапките на диференцијално пондерирање на тестовите на знаење и способности од типот повеќечлен избор се очекува да обезбедат подобра дискриминативност на применетите тестови, односно подобро разликување помеѓу испитаниците кои одговараат на тестот отколку што тоа го прават конвенционалните начини на бодување. Зголемената дискриминативност на диференцијално пондерираните тестови би довела до зголемена релијабилност и истовремено до подобра предиктивна валидност на применетите тестови, под претпоставка дека зголемената дискриминативност ги одразува вистинските разлики во знаењата и способностите на испитаниците и дека таквите разлики се во врска со упехот на критериумот.

Диференцијално пондерирање на ајтемите

За да се земе предвид различниот придонес на ајтемите во мерењето на варијаблата знаење или способност, на ајтемите им се доделува различен износ на бодови. Износот на бодовите се определува врз основа на тежината на ајтемите, било априористички, било емпириски. За да се обезбеди можност за дијагностицирање и мерење на различните степени на знаење, односно способност, применетиот метод на бодување подразбира припишување различни пондери на ајтемите зависно од нивното ниво на тежина. Во тој случај, текст-скорот на испитаниците претставува сума од постигнатите ајтем-скорови кои се пресметуваат по формулата: $Y_{ji} = w_j x_{ij}$

при што „Y_{ij}" е ајтем-скорот на индивидуата i на ајтемот j, „w_j" е пондерот припишан на ајтемот j, а „x_{ij}" е износот на бодови постигнат со конвенционално бодување (вредностите на x_{ij} можат да изнесуваат 1 за точно одговорените ајтеми и 0 за сите други ајтеми, или 1 за точно одговорените, 0 за испуштените и $-\frac{1}{k-1}$ за неточно одговорените ајтеми).

Највисоки се ајтем-скоровите за точно одговорените најтешки ајтеми, а најниски ајтем-скоровите или за неточно одговорените најтешки ајтеми или за неточно одговорените ајтеми без разлика дали се потешки или полесни. Оттука произлегува дека највисок тест-скор постигнуваат оние испитаници кои точно одговараат на најголемиот број потешки ајтеми, а најнизок тест-скор или оние кои неточно одговараат на најголемиот број од потешките ајтеми, или оние кои неточно одговараат на најголемиот број ајтеми.

Еден од првите обиди да се процени потребата од диференцијално пондерирање на ајтемите е направен во студијата на Douglass и Spencer уште во 1923 година. Авторите дале приказ на резултатите добиени со примена на четири стандардизирани теста на знаење по алгебра. За секој тест биле пресметувани по два вида скорови: едните базирани на пондерирање, другите без пондери. Двете серии на скорови за секој тест посебно биле корелирани и добиените Пирсонови коефициенти на корелација изнесувале 0,98, 0,99, 0,995 и 0,996. Иако Douglas и Spencer не извеле дефинитивен заклучок, резултатите кои ги добиле недвосмислено укажале на незнатна вредност на пондерирањето со оглед на тоа што најдените корелации помеѓу скоровите добиени со пондерирање и без пондерирање биле многу високи (над 0,95).

Друго истражување кое ја довело во прашање потребата од пондерирање на ајтемите било спроведено од West (1924). Истражувањето било поттикнато од прашањето дали скоровите на тестовите треба да претставуваат обичен збир од точно одговорените ајтеми, или треба да се пресметуваат врз база на пондерирање на ајтемите според нивната тежина. Авторот ги применил двата начина на бодување на два теста на сфаќање на усно прочитани текстови и добиените резултати ги споредил. Пондерите ги определил емпириски, врз основа на процентните вредности на придонес на секој ајтем во постигнувањето на целиот тест.

Коефициентите на корелација пресметани за скоровите добиени со пондерирање и за скоровите добиени конвенционално по продукт-момент методата, се покажале многу високи (сите преку 0,985). Ниеден од коефициентите не укажал на доволно ниска поврзаност за да се оправда заклучокот дека пондерирањето значајно го менува скорот или рангот на испитаниците.

Студијата на West опфаќа и обид да се процени вредноста на пондерирањето на ајтемите во тестовите на способности. За таа цел биле најдени коефициентите на корелација помеѓу скоровите пресметани по двата начина на бодување применети на резултатите од шест субтестови на Армискиот алфа тест. Сите корелации се покажале доста високи (коефициентите изнесувале од 0,932 до 0,984 — просечно околу 0,958), но сепак пониски од оние добиени на претходно применетите тестови. Ваквиот наод авторот го протолкувал како можен показател на две нешта: или

дека пондерирањето на помал број ајтеми во група (како што е тоа случај со субтестовите) им придава поголемо значење на пондерите, или дека определени тестови поставуваат поголемо барање за пондерирање (кое ги прави поадекватни мерни инструменти на специфичните или општата способност).

Во 1938 година Wilks направил аналитички студија на проблемот на диференцијално пондерирање на ајтемите. Студијата покажала дека со зголемување на бројот на ајтемите во тест составен од позитивно корелирани ајтеми, се зголемува и корелацијата помеѓу два по случајност пондерирани составни дела на истите ајтем-скорови. Во склад со овој наод авторот заклучил дека за долгите тестови со позитивно интеркорелирани ајтеми не е значајно како се пондерирани индивидуалните ајтеми, затоа што релативниот редослед на скоровите на испитаниците тежнее да остане непроменет кога се применуваат различни методи за пресметување на линеарно поврзаните скорови.

Stalnaker (1938) го истражувал пондерирањето на ајтемите на еден стандардизиран тест на знаење кој се користел при селекција на студенти. Ги корелирал скоровите базирани на одговорите пондерирани од страна на неколку стручњаци од областа, со скоровите добиени на истите одговори без пондерирање. Добиените коефициенти на корелација (над 0,97) покажале дека постои голема усогласеност помеѓу скоровите добиени со пондерирање и без пондерирање, наведувајќи на заклучок дека априористичкото пондерирање на тест-ајтемите не претставува никакво подобрување во однос на конвенционалниот начин на бодување.

Студиите на Douglas и Spencer, West, Wilks, Stalnaker и некои други обезбедиле основа за песимистичко гледање на потребата од диференцијално пондерирање на ајтемите. Надеж дека од пондерирањето на ајтемите може да се очекува извесно подобрување нудат единствено два приода— едниот понуден од Birnbaum (1968), а другиот од Cleary (1966).

Cleary понудила модел на мултипла регресија кој дозволува емпириско јавување на индивидуални разлики без никакви ограничувања од какви и да е априористички концепции. Моделот го тестирала со измислени и реални податоци за да покаже дека ефикасно ја редуцира варијансата на грешка во предвидувањето и дека обезбедува пондери кои се непроменливи за различни примероци испитаници и за различни видови предиктори. Моделот се базира на припишување разлачни сетови на регресивни пондери на секој испитаник посебно. Најголемата предност на овој модел е што нуди емпириски метод за процена дали предикцијата може да се подобри преку отстапување од вообичаениот модел на мултипла регресија и колку димензии се потребни за нејзино максимално подобрување. Припишувајќи различни сетови од пондери за различни испитаници, моделот на Cleary може да обезбеди начин за пондерирање на ајтем-скоровите кој

овозможува подобро предвидување на критериумите отколку што е можно со примена на ист сет пондери за сите испитаници.

Суштината на Birnbaum-овиот трипараметарски логистички модел на латентни особини (како што е прикажан во книгата на Lord и Novick, 1968) е во тоа што применува различни пондери не само за различни ајтеми, туку и за различни нивоа на способности. На секој ајтем му се припишува определен пондер на начин кој овозможува максимално зголемување на ефикасноста на тестот за секое однапред определено ниво на способности. Секој таков оптимален ајтем-пондер се определува како функција од нивото на способности за кое се бара максимална дискриминативност (пондерот за ајтемот j не е повеќе w_j, туку w_{jl} — „l" го означува нивото на способности).

Како резултат на применетиот систем на бодување се добива најголемо подобрување за скоровите на најмалку способните испитаници кои инаку одговараат на тешките ајтеми главно врз база на случајно погодување со што ја зголемуваат грешката во своите скорови. Изгледа дека методот на Birnbaum е во можност да ги поништи последиците од таквото погодување на тој начин што им припишува помали пондери на ајтемите кои се тешки за конкретното ниво на способности, а поголеми пондери на полесните ајтеми за тоа ниво на способности.

Lord (1968) го испробал овој модел на вербалниот дел на еден тест на способности (Scholastic Aptitude Test) применувајќи го на скоро 3.000 испитаници кои конкурирале за прием на американските универзитети. Тој истакнал дека моделот може со успех да се примени само на ајтеми на кои со конвенционално бодување можат да се добијат 1 или 0 бодови (за точен, односно неточен одговор) и на податоци кои не опфаќаат испуштени одговори. Резултатите од студијата на Lord одат во прилог на предложениот модел на Birnbaum, иако авторот предупрдува дека може да се случи заклучоците да не важат за случај кога има потреба истите ајтеми да се бодуваат по формула (со корекција за погодување).

Диференцијално пондерирање на алтернативите

Диференцијалното пондерирање на ајтем-алтернативите се базира на претпоставката дека нуди дополнителни информации за знаењето или способноста на испитаниците изразени во вид на варијанса која може да се припише на изборот помеѓу неточните алтернативи од страна на испитаниците кои не се во состојба да го идентификуваат точниот одговор. Според Davis (1959, стр. 292), мерењето на оваа варијанса со тестовите на знаење бара: (1) неточните алтернативи на секој ајтем да претставуваат различни степени (нивоа) на парцијално знаење и незнаење со оглед на мерената вирјабла и (2) припишувањето пондери на

секоја од неточните алтернативи да ја приближува алтернативата кон нејзиното вистинско место на континуумот кој се движи од парцијално знаење кое е одвај неадекватно за да се идентификува точниот одговор, преку различни степени на парцијално знаење и незнаење до потполно незнаење. До кој степен ќе се задоволи првото барање зависи од умешноста на составувачот на ајтемите. Второто барање се задоволува со определување пондери за секоја алтернатива на ајтемот по логички (априористички) или емпириски пат (врз основа на уделот на алтернативата во релијабилноста или валидноста на скоровите).

Диференцијалното пондерирање на ајтемите во тестовите на способности се базира на истата претпоставка, само што се бара неточните алтернативи да одразуваат различни нивои на отсуство на индикаторот кој се одразува со ајтемот, односно, конинуумот да се движи од одвај неадекватна способност за идентификување на точниот одговор до потполно отсуство на индикаторот на способноста која се мери со ајтемот.

Значи, диференцијалнто пондерирање на алтернативите се прави за да се овозможи квантитативна диференцијација помеѓу испитаниците кои, за даден ајтем, ја избираат неточната алтернатива која укажува на незнаење или погрешно знаење, односно на отсуство на индикаторот на способност за тестираната варијабла. Кога се користи постапка од овој вид, се доделуваат различни пондери на секоја понудена алтернатива за сите ајтеми и тест-скорот се претставува со сумата од пондерите на точните и неточните алтернативи кои испитаникот ги избрал одговарајќи на тестот.

Пондерирањето на алтернативите започнало дваесеттите години на овој век со работата на Strong во областа на професионалните интереси. Тој ги бодувал ајтем-алтернативите во неговиот инвентар на професионални интереси така да диференцираат различни групи на занимања на тој начин што ги пресметувал процентите на одговори дадени од страна на секоја група занимања на секоја понудена алтернатива и тие проценти ги користел за да определи како да ги оценува алтернативите.

Диференцијалното пондерирање на алернативите во областа на тестирање на знаењата и способностите меѓу првите го применил Guttman (1941). Тој предложил техника на емпириско пондерирање според која доделувањето пондери на секоја од алтернативите се базира на тест-скорот на испитаниците кои ја избрале таа алтернатива. Тој презентирал податоци за да покаже дека како пондер може да се користи аритметичката средина на тест-скоровите на оние испитаници кои се определиле за конкретната алтернатива.

Guttman-овиот модел на пондерирање може да се применува откако ќе се задоволат три услова: (а) аритметичката средина на тест-скоровите на испитаниците кои ја избрале точната алтер-

натива треба да биде значајно повисока од аритметичките средини на тест-скоровите на испитаниците кои избрале која и да е друга аллернатива; (б) помеѓу аритметичките средини на тест-скоровите на испитаниците кои ги избрале различните неточни алтернативи (вклучувајќи ги и испуштените избори) треба да постои значајна разлика; и (в) добиениот редослед на аритметичките средини да одговара на тоа колку се алтернативите неточни како што е проценето од страна на стручњаците од областа на мерената варијабла (алтернативите подалечни од точниот одговор да имаат пониски аритметички средини).

Davis и Fifer (1959) спровеле истражување кое имало за цел да ги определи ефектите од диференцијалното пондерирање на алтернативите врз релијабилноста и валидноста на скоровите добиени на два теста на аритметичко резонирање. Тестовите ги задале на голем, репрезентативен примерок испитаници земени од популацијата на која ѝ се тестовите наменети. Просечниот скор на критериум-варијаблата (просечниот скор на тестот) кој го постигнале испитаниците кои избрале определена алтернатива, претставувал пондер за таа алтернатива.

Бодувањето на тестовите кое ги земало предвид пондерите на избраните алтернативи довело до статистички значаен пораст на релијабилноста (од 0,68 на 0,76). Практичното значење на овој наод се гледа во тоа што порастот во релијабилноста на тест-скоровите бил постигнат без продолжување на должината на тестот, без зголемување на времето на задавањето на тестот, како и без намалување на конкурентната валидност. Всушност, со пондерираното бодување на тестот кој се состоел од 45 ајтеми се добиле скорови кои се релијабилни колку и скоровите кои би резултирале од конвенционално бодување на тест кој брои 67 ајтеми (на кои се доделува 1 бод за секој точен одговор, а 0 бодови за секој неточен).

Нешто подоцна, Davis (1959) предложил поедноставен начин за емпириско определување на пондерите за ајтем-алтернативите. Поаѓајќи од тоа дека директното пресметување на просечниот скор на критериумот за групата испитаници кои избираат одделна алтернатива во рамките на секој ајтем е прилично сложено и напорно, Davis предложил пондерите да се базираат на просечниот стандарден скор на критериумот постигнат од групата испитаници кои ја избрале конкретната алтернатива. Тој понудил табела во која се дадени проценети просечни стандардни скорови кои ги претставуваат оптималните пондери за секоја ајтем-алтернатива во тест од типот повеќечлен избор. До проценетите просечни стандардни скорови на испитаниците кои избрале определена алтернатива се доаѓа со посредство на пропорците на испитаници од најдобрата и најслабата група (кои ги сочинуваат по 27% најдобри односно најслаби испитаници на критериумот) кои ја избрале дадената алтернатива. Проценети-

те просечни стандардни скорови се изразени во вид на цели броеви кои се движат од —9 до +9. Како илустрација може да послужи долуприкажаниот фрагмент од табелата на Davis:

ВРЕДНОСТИ ОД ТАБЕЛАТА НА ПОНДЕРИ НА DAVIS

		пропорција на испитаници од добрата група кои ја избрале алтернативата				
		12	14	16	18	20
пропорција на испитаници од слабата група кои ја избрале алтернативата	01	8	8	8	9	9
	02	6	6	7	7	7
	04	4	4	5	5	5
	06	2	3	4	4	4
	08	1	2	2	3	3
	10	1	1	2	2	3

Авторот пружил доказ дека проценетите пондери прочитани од понудената табела (изразени како проценети просечни стандардни скорови на критериумот) по својата нумеричка вредност се многу слични со пондерите кои би се добиле со директно пресметување кое е многу сложено. Применувајќи ја понудената постапка на два паралелни теста на аритметичкото резонирање, Davis нашол висока корелација помеѓу проценетите пондери (0,64 за точните алтернативи и 0,67 за дистракторите), што му овозможило да заклучи дека пондерите определени според опишаната процедура се умерено стабилни (со зголемување на примерокот испитаници на кои се задава вака пондериранот тест се добива поголема стабилност на пондерите).

Sobers и White (1969) спровеле емпириско истражување на влијанието на пондерирањето на алтернативите врз релијабилноста и предиктивната валидност на скоровите. За определување на пондерите авторите ја користеле постапката предложена од Davis (1959), односно неовата табела на пондери. Како критериум за определување на најдобрите и најслабите 27% од испитаниците користени се скоровите на еден тест на знаење по алгебра (добиени како суми од точните одговори). Стандардизираниот тест по алгебра на кој било применувано диференцијално пондерирање го решавале четири групи испитаници.

Кога тестот бил бодуван конвенционално (со по 1 бод секој точен одговор и 0 бодови неточните одговори) добиените коефициенти на предиктивна валидност изнесувале: 0,767, 0,713, 0,745 и 0,674. Кога се применети добиените пондери, коефициентите на валидност се зголемиле за: 0,004, 0,000, 0,023 и 0,025. Коефициентите на релијабилност, пресметани по Спирман-Брауновиот метод, се промениле за 0,004 (од 0,891 добиено за конвен-

ционалната постапка), 0,015 (од 0,871), 0,006 (од 0,875) и 0,013 (од 0,883). Коментирајќи ги добиените резултати кои не одат во прилог на диференцијалното пондерирање, авторите предлагаат пондерирањето на алтернативите да се користи кај тестовите кои се состојат од ајтеми со најдобар одговор наместо од ајтеми со еден точен одговор.

Во својата студија, Hambleton, Roberts и Traub (1970) ги споредиле диференцијалното пондерирање на ајтем-алтернативите и тестирањето на увереноста во точноста на избраните алтернативи по однос на релијабилноста и валидноста. Како основа за споредување послужиле резултатите добиени со задавање на истиот тест на знаење под конвенционални услови (при што се земани предвид само точните одговори бодувани со 1 бод). Авторите очекувале тест-скоровите кои ќе се добијат со примена на двете експериментални постапки (диференцијалното пондерирање и тестирањето на увереноста) да содржат повеќе информации за состојбата на знаењето на испитаниците затоа што претпоставувале дека тие постапки овозможуваат да се земе предвид парцијалното знаење на испитаниците.

Постапките на диференцијално пондерирање опфаќале два различни вида пондери добиени априористички на два различни начина, користејќи рангирање на понудените алтернативи од страна на 22 експерта од дадената област (увод во психолошко и педагошко мерење). Според посложениот начин, пондерите за дистракторите се движеле во интервалот од —1,5 до +1,5 (највисоката вредност се однесувала на ди тракторите кои биле рангирани како најблиски до точниот одговор), а пондерите за точните алтернативи секогаш изнесувале +2. Според поедноставниот начин, алтернативата со најнизок просечен ранг (тоа е точниот одговор) добивала пондер 4, дистракторот со втор најнизок просечен ранг добивал пондер 3 и така натаму до дистракторот со највисок просечен ранг (алтернативата оценета како најдалечна од точниот одговор) кој добивал пондер 0.

Применетата постапка на тестирање на увереноста претставувала еден вид пробабилистичко тестирање кое барало од испитаникот да ја означи својата увереност во точноста на секоја од алтернативите на тој начин што ќе распореди 100 поени на алтернатирите кои ги смета за можен одговор на ајтемот. Ајтем-скоровите на испитаникот зависеле од изразеното количество на увереност во точноста на точната алтернатива, а неговиот тест-скор (S) се пресметувал по формулата:

$$S_i = \sum_{j=1}^{n} (1 + \log r_{ij}), \quad 0{,}01 \leqslant r_{ij} \leqslant 1{,}0$$

при што „r_{ij}“ е износот на увереноста во точноста на точната алтернатива на ајтемот j изразена од страна на индивидуата i.

За утврдување на релијабилноста бил применет пар-непар методот на делење на тестот на два дела. Валидноста пак, била проценувана со корелирање на добиените тест-скорови со скоровите на друг тест на знаење од истата област зададен како завршен писмен испит.

Резултатите од студијата покажале дека постапката на тестирање на увереноста произвела највалидни, но најмалку релијабилни скорови (коефициентот на валидност изнесувал 0,720, а коефициентот на релијабилност 0,655). Со конвенционалната постапка на тестирање се добиле најмалку валидни скорови (коефициентот на валидност изнесувал 0,621) кои се покажале исто релијабилни како и скоровите добиени со поедноставниот начин на диференцијално пондерирање (коеф. на релијабилност изнесувал 0,71). Со посложениот начин на диференцијално поднерирање се добил коефициент на релијабилност 0,692 и коефициент на валидност 0,703 (наспроти коефициентот на валидност 0,673 добиен за поедноставниот начин на диференцијало пондерирање).

Авторите понудиле и друг начин за спроведување на резултатите од релијабилноста и валидноста на применетите постапки, преку факторите на подобрување пресметани по формулите:

$$K_r = \frac{r_{kk}\ (1-r_{II})}{r_{II}\ (1-r_{kk})}$$

при што „Kr" е факторот на подобрување на релијабилноста за експерименталните постапки, „r_{kk}" е коефициентот на релијабилност на експерименталната постапка, а „r_{II}" е коефициентот на релијабилност за конвенционалната постапка;

$$K_v = \frac{v_{kk}^2\ (1-r_{II})}{v_{II}^2 - v_{kk}^2\ r_{II}}$$

каде „Kv" е факторот на подобрување на валидноста за експерименталните постапки, „v_{kk}" е коефициентот а валидноста на експерименталната постапка, а „r_{II}" е коефициентот на валидност на конвенционалната постапка.

Факторите на подобрување покажуваат за колку треба да се продолжи конвенционално применетиот тест за да се постигне истата релијабилност односно валидност како што е добиена со експерименталните постапки на задавање и бодување на истиот тест. Добиените фактори на подобрување на релијабилноста за двете постапки на диференцијално пондерирање на алтернативите изнесуваат 0,917 за посложената и 1,005 за поедноставната, додека за постапката на тестирање на увереноста 0,773. Факторите на подобрување на валидноста изнесувале 4,124 и 2,050 за диференцијалното пондерирање и 8,533 за тестирање на увереноста.

Иако резултатите од студијата наведуваат на заклучок дека и двете експериментални постапки довеле до подобрување на валидноста, ваквата констатација мора да се прифати со резерва главно од две причини. Прво, групите испитаници на кои се применувани постапките на тестирање биле толку мали (околу 70) што не дозволиле ниедна од најдените разлики во коефициентите на валидност да се покаже статистички значајна. Второ, за постапката на тестирање на увереноса било потребно 10 минути подолго време за задавање отколку за другите постапки, што може да значи дека со изедначување на времето за тестирање би дошло до намалување на валидноста на оваа постапка.

Hendrickson (1971) направил обид да утврди на кој начин пондерирањето на алтернативите влијае на интерната конзистентност на тестот и на интеркорелацијата помеѓу ајтемите, користејќи четири субтестови на еден тест за селекција и прием на студенти (Scholastic Aptitude Test-SAT). Во студијата се обработени податоците добиени на примерок од 10000 случајно избрани испитаници кои конкурирале за прием на американските универзитети.

Пондерите за секоја алтернатива биле определувани емпириски според постапката која ја предложил Guttman (1941), а која се базирала на тест-скоровите на испитаниците кои ја избрале дадената алтернатива. Испитаниците биле поделени во две групи — пондерите определени врз база на скоровите на едната група биле применувани при бодувањето на одговорите на другата група. На тој начин се обезбедила двојна крос-валидација.

Тест-скоровите добиени со диференцијално пондерирање биле спроведувани со тест-скоровите добиени конвенционално со бодување по формула. Коефициентите на интерна конзистентност на субтестовите на SAT покажале подобрување при користењето на Guttman-овите пондери (најдените коефициенти за скоровите добиени со пондерирање се движеле во интервалот 0,8119—0,9214, додека за конвенционално пресметаните скорови интервалот бил понизок: 0,7818—0,8711). Процентот на продолжување на конвенционално бодуваниот тест до обезбедување на иста релијабилност како при диференцијалното пондерирање се разликувал за различни испитаници и за различни субтестови. Глобалниот пораст на должинната на тестот се движел од 19% до 78%, со просечен износ од 49%. Битно е што ваков значаен пораст се добил без продолжување на времето за решавање на тестот.

Истовремено, оваа студија покажала дека во најголем број случаи постои обратнопропорционална поврзаност помеѓу интерната конзистентност на определен субтест и корелацијата на тој субтест со другите субтестови во рамките на целиот тест. Самото тоа што интерната конзистентност на субтестовите се зголемила без додавање на нови ајтеми, покажува дека ајтемите во субтестот станале похомогени. Ова пак, навестува дека факторската струк-

тура на субтестовите се променила и дека пондерираните субтестови се сосојат од помал број фактори отколку што е тоа случај со конвенционално бодуваните субтестови. Со други зборови, можно е високата интеркорелација помеѓу ајтемите во определен субтест да претставува индикатор дека овие ајтеми мерат помалку аспекти на способноста. Изгледа дека пондерирањето на алтернативите го променило она што тестот го мери.

Collet (1971) извршила експериментално споредување на релијабилноста и валидноста на скоровите добиени конвенционално, со пондерирање на ајтемите и со елиминационо тестирање. За добивање на скоровите на конвенционален начин користено е бодување по формула. Применетата постапка на диференцијално пондерирање на ајтемите им припишувала +1 бод на точната алтернатива и +0,5, 0, —0,5 или —1 бод на дистракторите. Пондерите кои им биле припишувани на дистракторите се определувани со линеарна трансформација на сировите пондери добиени како аритметички средини на сумите точни одговори на тестот за испитаниците кои ја избрале дадената алтернатива. Постапката на елиминационото тестирање барала од испитаниците да одговараат со избирање и означување на неточните понудени алтернативи. Скорот на секој ајтем бил претставен со бројот на означените неточни алтернативи намален за к—1 ако била означена и точната алтернатива меѓу неточните.

Резултатите од студијата ја поткрепиле претпоставената супериорност на елиминационото тестирање. Добиените коефициенти на релијабилност изнесувале 0,858 за елиминационото тестирање, 0,809 за конвенционалното бодување и 0,725 за диференцијалното пондерирање. Коефициентите на критериум-валидност изнесувале 0,777 за елиминационото тестирање, 0,668 за конвенционалното бодување и 0,646 за диференцијалното пондерирање. Резултатите од тестирањето на значајноста укажале на супериорност не елиминационото тестирање во три од четири споредувања. Скоровите добиени со елиминационото тестирање се покажале порелијабилни од скоровите добиени со пондерирање на алтернативите и повалидни од другите два вида скорови. Истовремено, помеѓу скоровите добиени конвенционално и со диференцијално пондерирање не се покажала статистички значајна разлика ниту во поглед на релијабилноста, ниту во поглед на валидноста.

Patnak и Traub (1973) спровеле истражување кое имало за цел да ги спореди конвенционалните начини на бодување и диференцијалното пондерирање во поглед на релијабилноста и предиктивната валидност. Во истражувањето бил применет групен тест на општа интелигенција кој се состоел од 69 ајтеми со по 5 понудени алтернативи.

Пондерите за алтернативите биле изведувани априористички, врз основа на степенот на точност на секоја алтернатива, за секој ајтем посебно, како што било проценето од страна на 61 проценувач. Најнискиот пондер доделуван на најдалечниот дистрактор

секогаш имал вредност 0, за да не се случи да носи понеќе бодови отколку кога не е даван одговор. Максималниот пондер за секој ајтем не бил ограничен — неговата вредност зависела од извршеното рангирање од страна на проценувачите. Највисокиот доделен максимален пондер изнесувал 3,994, а најнискиот 0,820.

Пресметаните коефициенти на релијабилност по пар-непар методот на делење на тестот на два дела изнесувале: 0,894 за скоровите добиени како сама од точните одговори, 0,881 за скоровите добиени со бодување по формула и 0,915 за пондерираните скорови, укажувајќи на највисока релијабилност на скоровите добиени со диференцијално пондерирање на алтернативите. Разликите помеѓу коефициентот на релијабилност за скоровите добиени со пондерирање и коефициентите на релијабилност за двата вида конвенционално добиени скорови се покажале значајни, додека разликата помеѓу коефициентите на релијабилност за едните и за другите конвенционално пресметани скорови не била значајна.

Коефициентите на валидност биле пресметани како коефициенти на корелација помеѓу оценките на испитаниците дадени од страна на нивните наставници и оценките на испитаниците добиени врз основа на резултатите од трите применети постапки на бодување. На тој начин била најдена највисока валидност за скоровите добиени со примена на бодувањето по формула (0,4314), а најниска валидност за скоровите добиени со диференцијално пондерирање (0,4037). Разликата помеѓу коефициентите на валидност за едниот и за другиот конвенционален начин на бодување не се покажала значајна, но, затоа пак, разликите помеѓу коефициентите на валидност за конвенционалните постапки и коефициентот на валидност за пондерирањето на алтернативите биле значајни на ниво 0,01.

Една друга студија во која се проучува релијабилноста и валидноста на диференцијалното пондерирање на алтернативите спроведена е од Kansup и Hakstian (1975). Авторите примениле две форми на тест на математичко резонирање и две форми на тест на вербални способности на голем примерок испитаници. Одговорите биле бодувани на два начина: конвенционално (тест-скорот бил претставен со бројот на точно одговорени ајтеми) и со пондерирање на алтернативите (тест-скорот претставувал сума од пондерите на алтернативите кои испитаникот ги избрал).

Пондерите припишувани на ајтем-алтернативите биле утврдувани по логички пат — со рангирање на дистракторите од страна на 44 студенти кои доделувале вредност 4 на најблискиот дистрактор, вредност 3 на следниот најблизок итн., до вредност 1 на најдалечниот дистрактор. Аритметичките средини на доделените вредности за секој дистрактор ги претставувале бараните пондери. Пондерот за точната алтернатива изнесува 5.

Релијабилноста била изразувана преку алфа-коефициентот како мерка на интерната конзистентност на тестот и преку тест-ретест коефициентот како мерка на стабилност на скоровите. Валидноста била проценувана преку утврдување на корелација со три критериума: крајните оценки од сродните предмети, просечниот школски успех и вкупните скорови на двете форми на тестовите.

Студијата укажала на два главни наода во поглед на релијабилноста на постапката на диференцијалното пондерирање. Логичкото пондерирање довело до извесно зголемување на интерната конзистентност на скоровите на двата вида тестови (просечно околу 0,08). Меѓутоа, забележано е опаѓање на тест-ретест релијабилноста добиена со постапката на пондерирање (просечно околу 0,07), иако разликите во коефициентите не достигнале дури ни 0,10 ниво на значајност.

Резултатите од проценатa на валидноста покажуваат дека не постои статистички значајна разлика помеѓу скоровите на тестот на вербални способности добиени конвенционално и со пондерирање за ниеден од применетите критериуми. Истовремено, за скоровите на тестот на математичко резонирање добиени со пондерирањето најдено е значајно опаѓање на валидноста за сите применети критериуми.

Резултатите од студијата наведуваат на заклучок дека со диференцијално пондерирање не се доаѓа до суштествени дополнителни информации во врска со потенцијалите за постигнување на испитаниците кога пондерите се определуваат по логички пат. Со други зборови, студијата не ја оправдала потребата од примена на сложената и долга постапка на диференцијално пондерирање на ајтемалтернативите според априористички утврдени степени на точност.

Echternacht (1976) спровел истражување со цел да ја определи ефикасноста од примената на различни бодовни шеми, посебно на емпиристичките и априористичките начини на пондерирање на алтернативите наспроти конвенционалните начини на бодување и некои нивни модификации. Истражувањето било спроведено на огромен број испитаници (шест случајно избрани примероци кои броеле по 1000 субјекти) на кои им бил задаван посебно конструиран тест на способности во рамките на тестовите наменети за селекција и прием на постдипломски студии (General Record Examinations). Како примарна мерка на ефикасноста на применетите бодовни шеми користена е релијабилноста на добиените скорови изразена преку коефициентот алфа. Скоровите биле споредувани и во поглед на конкурентната валидност при што, како критериум се користени скоровите добиени на сличен, подолг тест.

Во студијата се споредувани скорови добиени со посредство на следните бодовни шеми:

(1) скорот како број на точно одговорени ајтеми: точните одговори се бодуваат со 1 бод, а неточните и испуштените со 0 бодови;

(2) бодување по формула: точните одговори се бодуваат со 1 бод, неточните со $-\frac{1}{k-1}$ а испуштените со 0 бодови;

(3) априористичко пондерирање на алтернативите: на точната алтернатива ѝ се доделуваат 6 бодови, на поблискиот дистрактор 1 бон, а на подалечниот —4 бодови (за испуштените одговори не се добивале поени);

(4) емпириско пондерирање на алтернативите: пондерите се добивале на тој начин што, прво, се бодувал тестот конвенционално (бодување по формула), а потоа, на секоја алтернатива ѝ се придавал пондер определен од аритметичката средина на стандардниот скор добиен на резултатите од преостанатите ајтеми за сите испаници кои ја избрале таа алтернатива (пондерите се пресметувале со примена на тестот на посебен примерок испитаници);

(5) Zl-бодовен систем (по Zinger, 1972): Точниот одговор се наградувал со 1 бод, испуштениот со 0 бодови, а неточниот со —c бодови пресметани по формулата:

$$c = \sum_{j=1}^{k-1} n_j^{\,2} \Big/ \left(\sum_{j=1}^{k-1} n_j \right)^2$$

каде „k" е бројот на алтернативите, „n_j " бројот на испитаниците кои одговарале на j-тиот дистрактор;

(6) Z2-бодовен систем (по Zinger, 1972): точниот одговор се наградувал со 1+b бодови, испуштениот со 0, а неточниот со —c бодови, при што

$$b = \sum_{j=1}^{k-1} (n_j - \bar{n})^2 \Big/ \left(n_t \sum_{j=1}^{k-1} n_j \right)$$

каде „n_t " го означува бројот на испитаници кои одговориле точно и

$$\bar{n} = \sum_{j=1}^{k-1} n_j \,/\, (k-1)$$

Според добиените коефициенти на релијабилност, емпирискиот систем на диференцијално пондерирање на алтернативите се покажал како најефикасен бодовен систем (коефициентот на

релијабилност изнесувал 0,85 во просек). Пораста во релијабилноста која е резултат на емпириско пондерирање одговара на 33 процентно просечно зголемување на должината на тестот кој се бодува по формула (коефициентот на релијабилност за бодувањето по формула изнесува 0,81 во просек), и на нешто помало зголемување кога се собираат само точно одговорените ајтеми (просечниот коефициент на релијабилност е 0,82). Истовремено, априористичкото кондерирање заедно со Z1 и Z2 системите на бодување не ја достигнало ни релијабилноста на бодувањето по формула (просените коефициенти на релијабилност за трите бодовни системи изнесувале 0,79, 0,80 и 0,79). Резултатите од проценатa на валидноста сосема одговараат на опишаните резултати од проценатa на релијабилноста. Највалидна се покажала постапката на емпириско пондерирање (просечниот коефициент на валидност изнесувал 0,861), наспроти валидноста на двете конвенционални постапки на бодување (ч ии коефициенти на валидност изнесувале 0,848 во просек). Најниска валидност се добила за априористичкото пондерирање и Z1 и Z2 системите на бодување (со просечни коефициенти на валидност 0,840, 0,843 и 0,840).

Имајќи ги предвид наведените резултати, авторот заклучил дека априористичкото пондерирање на алтернативите е помалку вредно од емпириското, па дури и од конвенционалните начини на бодување. Според авторот, единствено корисно е емпириското пондерирање затоа што овозможува да се сочува релијабилноста на тестот и кога се намалува неговата должина, што како практична последица има намалување на трошоците врзани за развивање ајтеми во тестот.

ЗАКЛУЧОК

Диференцијалното пондерирање на ајтемите и ајтем-алтернативите се јавило како обид за попрецизно изразување на степенот на поседување на варијаблата која се мери со тестот. Поаѓајќи од претпоставката дека припишувањето различни бодовни вредности на ајтемите или ајтем-алтернативите во тестот поверно ја одразува вредноста на ајтемот и алтернативата како индикатор на мерената варијабла, од диференцијалното пондерирање се очекува да биде поефикасна метода за процена на нивото на знаење или способности.

Еден начин за процена на ефикасноста на постапките на диференцијалното пондерирање е утвдување на корелација помеѓу скоровите добиени со пондерирање и скоровите добиени на конвенционален начин, со доделување ист број бодови на ајтемите, односно дистракторите. Придонесот на диференцијалното пондерирање се утврдува и преку споредување на релијабилноста

и валидноста на тест-скоровите добиени со доделување различни пондери на ајтемите или алтернативите со релијабилноста и валидноста на тест-скоровите добиени со конвенционално бодување.

Истражувањата на постапките на диференцијално пондерирање на ајтемите покажуваат дека користењето на пондерите ретко придонесува за зголемување на ефикасноста на системот на бодување. Најдените коефициенти на корелација помеѓу тест-скоровите добиени со доделување различни бодовни вредности на ајтемите и тест-скоровите добиени со доделување исти бодовни вредности на ајтемите, се покажале премногу високи, наведувајќи на заклучок дека диференцијалното пондерирање на ајтемите не претставува подобрување во однос на конвенционалните начини на бодување. Единствено постапките кои доделуваат на ајтемите диференцијални пондери кои не се исти за сите испитаници, внесуваат извесно подобрување во поглед на релијабилноста и валидноста на добиените тест-скорови. Резултатите од примената на Birnbaum-овиот модел на диференцијално пондирање зависно од нивоите на способности на испитаниците и од Clearyевата постапка на обезбедување регресивни пондери кои вариираат од еден до друг испитаник, би можеле да ја подобрат успешноста на методата на диференцијално пондерирање на ајтемите.

Подложени на емпириски проверки, и постапките на диференцијално пондерирање на алтернативите не се докажале како поуспешни техники за процена на нивото на знаење или способности на испитаниците од конвенционалните техники. Резултатите од емпириските истражувања се покажале доста противречни, не овозможувајќи да се тврди дека припишувањето различни пондери на ајтем-алтернативите ја зголемува ефикасноста на така добиените скорови. Веројатно може да се констатира дека добиените резултати изразени преку коефициентите на релијабилност и валидност укажуваат на извесна предност на постапките на емпириско определување на пондерите за алтернативите над постапките кои користат априористички начин на утврдување на пондерите.

Може да се резимира: емпириските наоди не се едногласни во повторување на очекуваните предности на постапките на диференцијално пондерирање определени според теориските образложенија за примена на овие постапки. Од една страна, тоа може да послужи како поттик да се вложат поголеми напори во развивање и усовршување на постапките на диференцијално пондерирање со цел да се докаже очекуваното подобрување во однос на конвенционалните методи на бодување. Од друга страна, пак, може да се искористи како препорака тест-скорот да се пресметува со сумирање на ајтем-скоровите добиени без пондерирање и третирани на конвенционален начин: сите ајтеми се еднакво

вредни индикатори на присуството на мерената варијабла, односно сите дистрактори се еднакво вредни индикатори на отсуството на мерената варијабла.

ЛИТЕРАТУРА

Guttman L. THE QUANTIFICATION OF A CLASS OF ATRIBUTES: A THEORY AND A METHOD OF SCALE CONSTRUCTION. Vo P. Horst, Predicton of personal adjustment. New York: Social Science Research Council, 1941, 319—348.

Davis F. B. ESTIMATION AND USE OF SCORING WEIGHTS FOR EACH CHOICE IN MULTIPLE-CHOICE ITEM. Educational and Psychological Measurement, 1959, 19, 291—298.

Dvis F. F. i Fifer G. THE EFFECT ON TEST RELIABILITY AND VALIDITY OF SCORING APTITUDE AND ACHIEVEMENT TESTS WITH WEIGHTS FOR EVERY CHOICE. Educational and Psychological Measurement, 1959, 19, 159—170.

Douglas P. L. i Spencer R. L. IS IT NECESSARY TO WEIGHT EXERCISES IN STANDARD TESTS? Journal of Educational Psychology, 1923, 14, 109—112.

Echternacht G. RELIABILITY AND VALIDITY OF ITEM OPTION WEIGHTING SCHEMES. Educational and Psychological Measurement, 1976, 36, 301—310.

Zinger A. A. A NOTE ON MULTIPLE-CHOICE ITEMS. Journal of American Statistical Association, 1972, 67, 340—341.

Kansup W. i Hakstian A. R. A COMPARISON OF SEVERAL METHODS OF ASSESSING PARTIAL KNOWLEDGE IN MULTIPLE-CHOICE TESTS: I. SCORING PROCEDURES. Journal of Educational Measurement, 1975, 12, 219—229.

Lord F. M. AN ANALYSIS OF THE VERBAL SCHOLASTIC APTITUDE TEST USING BIRNBAUM'S THREE-PARAMETER LOGISTIC MODEL. Educational and Psychological Measurement, 1968, 28, 989—1020.

Lord F. M. i Novick M. R. STATISTICAL THEORIES OF MENTAL TEST SCORES. Reading, Mass.: Addison-Wesley, 1968.

Patnaik D. i Traub R. E. DIFFERENTIAL WEIGHTING BY JUDGED DEGREE OF CORRECTNESS. Journal of Educational Measurement, 1973, 10, 281—286.

Petroska V. PROBABILITY SCORING: CONFIDENCE TESTING AND DIFFERENTIAL WEIGHTING. Magisterski trud, Teachers College, Columbia University, 1982.

Sobers D. L. i White, G. W. THE EFFECT OF DIFFERENTIAL WEIGHTING OF INDIVIDUAL ITEM RESPONSES ON THE PREDICTIVE VALIDITY AND RELIABILITY OF AN APTITUDE TEST. Journal of Educational Measurement, 1969, 6, 93—96.

Stalnaker J. M. WEIGHTING QUESTIONS IN THE ESSAY-TYPE EXAMINATION. Journal of Educational Psychology, 1938, 29, 481—490.

Strong E. K. VOCATIONAL INTERESTS FOR MEN AND WOMEN. Stanford, Calif.: Stanford University Press, 1943.

Hambleton R. K., Roberts D. M. i Trab R. E. A COMPERISON OF THE RELIABILITY AND VALIDITY OF TWO METHODS FOR ASSESSING PARTIAL KNOWLEDGE ON A MULTIPLE-CHOICE TEST. Journal of Educational Measurement, 1970, 7, 75—81.

Hendrickson G. F. THE EFFECT OF DIFFERENTIAL OPTION WEIGHTING ON MULTIPLE-CHOICE OBJECTIVE TESTS. Journal of Educational Measurement, 1971, 8, 291—296.

Cleary A. AN INDIVIDUAL DIFFERENCES MODEL FOR MULTIPLE REGRESSION. Psychometrika, 1966, 31, 215—224.

Collet L. S. ELIMINATION SCORING: AN EMPIRICAL EVALUATION. Journal of Educational Measurement, 1971, 8, 209—214.

West R. V. THE SIGNIFICANCE OF WEIGHTED SCORES. Journal of Educatonal Psychology, 1924, 15, 302—308.

Wilks S. S. WEIGHTING SYSTEMS FOR LINEAR FUNCTIONS OF CORRELATED VARIABLES WHEN THERE IS NO DEPENDENT VARIABLE. Psychometrika, 1938, 3, 23—40.

Violeta PETROSKA — BEŠKA

DIFFERENTIAL WEIGHTING WITH MULTIPLE — CHOICE TESTS

(S u m m a r y)

This study is a theoretical survey of the differential weighting procedures applied to multiple-choice ability tests. It discusses the differences between conventional scoring techniques and differential weigting models as special ways of scoring the examinees' responses. It reviews many proposed and tested models of assigning differential weights to each item (known as weighting test items) and to each response alternative of an item (known as weighting item options). It also presents the difference between logical and empirical ways of assigning weights to items or item alternatives.

Even though differential weighting procedures emerged as an attempt to provide more efficient method for measuring the examinees' ability level, this study has shown that the expectations should not be doublessly accepted. The evaluation of the contribution of differential weighting through comparing the reliability and validity of the obtained weighted item scores and weighted option scores with the conventionally obtained test scores has not proved the assumed advanteges completely, seading to the conclusion that most of the differential weighting methods are not worthy the extended time and effort which they require.