ISSN 1392-5016. ACTA PAEDAGOGICA VILNENSIA 2001 8

Matematikos testų lygiagrečių variantų ekvivalentumo problema

Gediminas Trakas

Vilniaus universitetas

Straipsnyje svarstomos Lietuvos mokyklose naudojamų objektyvių matematikos testų lygiagrečių variantų ekvivalentiškumo problemos. Aptariami keturi skirtingi lygiagrečių testo variantų sudarymo būdai, taip pat analizuojami trys konkretūs testai, kurie jau buvo naudoti 1997–1998 m. per bandomuosius matematikos brandos bei stojamuosius egzaminus. Kartu pateikiamos kai kurios rekomendacijos, kaip sušvelninti testo variantų skirtumą, ir galimybės išvengti netinkamo testų variantų naudojimo egzaminų praktikoje.

Temos aktualumas. Testas, kaip objektyvi respondento žinių ir gebėjimų įvertinimo priemonė, naudojamas įvairių egzaminų ar diagnostinių tyrimų metu. Dažnai egzamino metu pateikiami keli testo užduočių variantai. Toks skirtingų testo variantų (testo formų) parengimas ir naudojimas vienu metu dažniausiai taikomas dviem tikslais. Pirmas tikslas – tai korektiškumo (sąžiningumo) siekimas egzamino metu, t. y. nusirašinėjimo galimybių mažinimas. Antras tikslas – siekti, kad diagnostinio testavimo metu kuo daugiau užduočių atskleistų kiek galima didesnę dalį dalyko turinio.

Per pastaruosius kelerius metus keitėsi egzaminų turinys, forma ir administravimo principai, tačiau sąžiningumo problema lieka aktuali. Mažinti galimybes nusirašyti ypač svarbu per egzaminus, kurie vienaip ar kitaip gali lemti juos laikančiųjų ateitį (brandos, stojamųjų ar pan.). Siekdami mažinti testo nesąžiningo atlikimo aplinkybes, testų rengėjai dažnai ruošia vienam egzaminui keletą panašių testo variantų ir juos naudoja vienu metu. Taigi greta sėdintys egzaminuojamieji tuo pačiu metu atlieka skirtingas užduotis. Panašus poreikis atsiranda ir kai dėl didelio egzaminuojamųjų skaičiaus egzaminas vyksta keliais srautais paeiliui. Siekiant išvengti galimo informacijos nutekėjimo po pirmos grupės testavimo antrai grupei (kuri testuojama vėliau), naudojamos skirtingos ekvivalenčios užduotys.

Daugelis užsienio švietimo specialistų pripažįsta, kad naudojant skirtingus testo variantus labai sumažėja nusirašinėjimo galimybės. R. C. Hollinger ir L. Lanza-Kaduce aprašo tyrimą, kuriame moksleivių buvo prašoma įvertinti 20 skirtingų nusirašinėjimo prevencijos strategijų (R. C. Hollinger ir L. Lanza-Kaduce, 1996). Geriausiai įvertintas metodas, kurį naudojant patys testo užduviniai/klausimai yra sumaišomi ir taip sukurti „naujų" užduoties variantai pateikiami egzaminuojamiesiems. Tokį metodą net 82 proc. moksleivių įvertino kaip labai efektyvų arba efektyvų

Metodas, kai variantai sudaromi iš visiškai skirtingų uždavinių / klausimų, buvo minimas tik penkiose sąrašo vietose. Jį kaip labai efektyvų arba efektyvų rinkosi 67 proc. respondentų. Kitos strategijos, apie kurias buvo klausta, nebuvo susijusios su lygiagrečiais testo užduočių variantais.

Diagnostiniais testais paprastai siekiama ne įvertinti respondento gebėjimus, bet patikrinti mokomojo dalyko turinio suvokimo problemas ir pateikti praktinius patarimus pedagogams. Tuo atveju testavimo turinys yra platesnis ir jį perteikti testo klausimais reikia labai daug užduočių. Jas visas įtraukus į vieną testą užduotį, testavimas užtruktų ir būtų varginantis egzaminuojamajam. Tai, savaime aišku, atsilieptų testo rezultatų patikimumui. Dažniausiai tokiu atveju rengiami keli testų variantai, kuriuose ir padalijamas tas užduočių kiekis. O administruojant visus testo variantus vienu metu, gaunami įverčiai apie visos testuojamųjų grupės žinias ar gebėjimus testuojamojo turinio atžvilgiu.

Abiem minėtais atvejais testų variantai akivaizdžiai yra naudingi. Tačiau jų naudojimas kelia problemų. Pirmoji – skirtingų testo variantų ekvivalentumas. Ar parengti variantai yra vienodo sunkumo, turinio, panašios kitos savybės, kurios vienaip ar kitaip gali sudaryti skirtingas sąlygas juos laikantiesiems? Ar vieno testo skirtingų variantų rezultatus galime lyginti, naudoti juos respondentų žinių ir gebėjimų „kiekiui" įvertinti ar juo grindžiamai atrankai? Kiek variantai gali skirtis, kad jų rezultatai būtų palyginami, t. y. įvertintų tas pačias respondentų žinias ar gebėjimus ir tais pačiais įverčių vienetais? Ar būtų pagrįstas gavusio jį netenkinantį įvertinimą asmens skundas, nurodantis vienintelę priežastį – egzamino metu gavau sunkesnį X variantą, o Y variantas buvo lengvesnis?

Lygiagrečių testo variantų ekvivalentumas yra apibrėžiamas ne tik užduočių turinio lygiavertiškumu. Skirtingų lygiagrečių variantų rezultatai (įverčiai) turi atitikti tą patį testuojamų žinių ar gebėjimų lygį. Ką reikia daryti, kai iš anksto negalima numatyti, ar variantai bus vienodo sunkumo, ar kitos svarbios statistinės jų charakteristikos bus vienodos? Šią problemą testų teorijos specialistai pradėjo spręsti aštuntajame XX a. dešimtmetyje, kai skirtingų testo variantų taikymas vienu metu buvo gana dažnai praktikuojamas. Kita priežastis, lėmusi aktyvų šios problemos sprendimų ieškojimą, buvo testavimo kritikų klausimai ir visuomenės reikalavimai užtikrinti vienodas sąlygas visiems dalyvaujantiems testuojant respondentams. Pirmiausiai buvo pradėtos kurti testų variantų palyginimo procedūros, kurios rėmėsi skirtingomis testavimo teorijomis. Variantų sulyginimo sąvokos ir procedūros pirmą kartą išsamiai pateiktos W. A. Angoffo 1971 metais išleistoje publikacijoje (Angoff W. A., 1971). Vėliau šiai tematikai buvo skirti P. W. Hollando ir D. B. Rubino (Holland P. W., Rubin D. B., 1982) bei M. J. Koleno ir R. L. Brennano (Kolen M. J., Brennan R. L., 1995) darbai. Tradicine testų teorija grindžiamų testo variantų sulygimo metodai 1988 metais buvo pateikti Koleno, kaip atskiro testų teorijos mokymo modulio metodinė medžiaga (Kolen M. J., 1988). 1991 metais pasirodė analogiška L. L. Cooko ir D. R. Eignoro parengta metodinė medžiaga (Cook L. L., Eignor D. R., 1991) apie metodus, kurie remiasi moderniąja testų teorija, t. y. viena pažangiausių jos šakų – IRT teorija (angl. *Item Response Theory*). Be minėtų monografijų, šia tema publikuojama daug mokslinių straipsnių, tyrimų ataskaitų ir konferencijų pranešimų, susijusių su šia tematika. Šioms problemoms tirti skirti specialūs žurnalų *Applied Psychological Measurement* (1987) bei

Applied Measurement in Education (1990) leidimai.

Testų teorijos naudojimas vertinti žinias ir gebėjimus, testams kurti Lietuvoje yra gana nauja tema ir todėl informacijos apie testų variantų problemas, galimus jų sprendimo metodus neradome.

Tyrimo objektas – Lietuvos mokyklose naudojamų objektyvių matematikos testų lygiagrečių variantų ekvivalentumo problemos.

Jo tikslai: a) išnagrinėti lygiagrečių variantų naudojimą; b) statistiniais metodais palyginti lygiagrečius testus, nustatyti juos skiriančius veiksnius bei ištirti galimybes mažinti atsitiktinius ir sisteminius veiksnius, lemiančius variantų skirtingumo įtaką moksleivio rezultatams.

Tyrimų metodika. Atliekant tyrimą teko susipažinti su įvairiais lygiagrečių testo variantų sudarymo būdais. Pavyzdžiais iliustruosime keletą dažniausiai naudojamų.

1. Skirtingi testo variantai sudaromi naudojant panašius, ekvivalenčios tematikos uždavinius ar klausimus. Tai gali būti vardažodžių pakeitimas, skirtingų situacijų (konteksto) naudojimas sprendžiant uždavinius, kuriems reikia taikyti tas pačias žinias ar gebėjimus bei procedūras. Paprastai už tokius uždavinius / klausimus skiriamas vienodas taškų skaičius, ir jie būna santykiškai toje pačioje testo užduoties vietoje. Pavyzdys – 1998 metų Lietuvos bendrojo lavinimo mokyklos geografijos brandos egzamino I ir II variantų klausimai:

I variantas		II variantas	
Priskirkite Europos miestams jų įžymybes (nurodykite rodykle):		*Priskirkite Europos miestams jų įžymybes (nurodykite rodykle):*	
Berlynas	Luvras	Praha	Big Benas
Roma	Brandenburgo vartai	Londonas	Partenonas
Paryžius	Koliziejus	Atėnai	Karolio tiltas

2. Skirtingų testo variantų ekvivalentūs uždaviniai skiriasi skaitinėmis reikšmėmis. Šis būdas dažnai taikomas matematikos testuose. Matematikos uždavinių specifika leidžia sudaryti panašius uždavinius keičiant tik skaitinius koeficientus lygtyse, nelygybėse, probleminiuose uždaviniuose ir pan. Pavyzdys – 1998 metų Lietuvos bendrojo lavinimo mokyklos matematikos B lygio brandos egzamino uždaviniai:

I variantas	II variantas
Kas daugiau $f(1)$ ar $f'(1)$, kai $f(x)=\frac{x+3}{x-2}$?	Kas daugiau $f(1)$ ar $f'(1)$, kai $f(x)=\frac{x+1}{x-2}$?

3. Kai užduotys sudarytos iš klausimų, turinčių pasirenkamus atsakymus, skirtinguose testo variantuose to pačio klausimo atsakymų tvarka gali būti sukeista. Toks būdas visiškai nekeičia uždavinio ar klausimo turinio ir labai tinka, siekiant neleisti teisingus atsakymus nusirašyti nuo greta sėdinčių egzaminuojamųjų. Pavyzdys – 1997 metų VDU stojamojo matematikos testo variantai:

I variantas	II variantas
Kiek šaknų turi lygtis $\lvert x-1\rvert = x-3$? A) *šaknų neturi* B) 1 C) 2 D) 4 E) *be galo daug*	Kiek šaknų turi lygtis $\lvert x-1\rvert = x-3$? B) 1 B) 2 C) 4 D) *be galo daug* E) *šaknų neturi*

4. Kartais skirtingų variantų užduočių uždaviniai / klausimai tie patys, bet pateikiami ne ta pačia tvarka. Pavyzdžiui, pirmieji vieno varianto klausimai pateikiami kito varianto užduoties pabaigoje. Kaip pavyzdį būtų galima nurodyti tarptautinį TIMSS tyrimą (Čekanavičius V. ir kt., 1997). Jo metu testuojant Lietuvos moksleivius buvo naudoti 8 testo variantai. Visų variantų uždaviniai buvo sugrupuoti (žymint nuo A iki Z). Visos grupės kituose testo variantuose buvo naudojamos skirtingose vietose, kaip parodyta 1 lentelėje:

1 lentelė. ***Atliekant TIMSS tyrimą naudotų uždavinių grupių vieta testuose***

Grupės vieta teste	Testo variantas							
	1	2	3	4	5	6	7	8
1	B	C	D	E	F	G	H	B
2	A	A	A	A	A	A	A	A
3	C	D	E	F	G	H	B	Q
4	S	W	T	X	U	Y	V	–
5	E	F	G	H	B	C	D	R
6	I	J	K	L	M	N	O	P
7	T	X	U	Y	V	Z	W	–

Taip sudarius variantus visi uždaviniai, išskyrus A grupės, atsidūrė skirtingose testų vietose. Kai kurios grupės buvo įdėtos į tris skirtingus variantus, kai kurios į du, o grupės nuo I iki P skirtinguose variantuose nesikartojo. Taigi moksleiviai vienu metu sprendė skirtingų variantų nevienodus uždavinius. Skirtingų variantų ekvivalentumas čia buvo užtikrinamas įdedant panašaus sunkumo, bet ne tapataus turinio uždavinius. Šis testas buvo naudojamas diagnostiniais tikslais ir kiekvieną uždavinį sprendė ne mažiau nei 300 atsitiktinai atrinktų moksleivių, o kai kuriuos ir daugiau nei 900. Šios sąlygos leidžia statistiškai patikimai įvertinti moksleivių, jų grupių ar visos generalinės aibės testuojamus gebėjimus ir žinias.

Tiriant naudotų testų ir respondentų duomenų bazių aprašymas

Šiame tyrime nagrinėti trys matematikos testai, naudoti 1997–1998 metais. Šie testai buvo rengti ir administruoti keliais lygiagrečiais variantais. Kiekvieno testo variantų sudarymo principai buvo skirtingi, tačiau bendras jų visų

naudojimo tikslas – sumažinti nusirašinėjimo galimybes. Testų variantai buvo sudaromi ir siekiant kuo didesnio jų turinio panašumo.

Pirmas matematikos testas buvo naudotas 1998 metais bandomojo matematikos A tipo (turiniu atitinkančio bendrojo lavinimo mokyklos A lygio mokymo programą) brandos egzamino metu. Kiekvieną testo variantą sudaro 20 skirtingų uždavinių, iš kurių 10 turi pasirenkamuosius atsakymų variantus, į tris reikia tik trumpai atsakyti bei septyni sudėtiniai uždaviniai, į kuriuos reikia tarpinių atsakymų ir visiškai juos išspręsti. Testas buvo administruojamas Panevėžio apskrityje ir jį laikė 879 vienuoliktų klasių moksleiviai 32 mokyklose. Analizei atsitiktinai atrinkti 244 pirmo ir 246 antro testo varianto darbai.

Šie testo variantai buvo sudaryti iš tų pačių uždavinių pakeičiant skaitinius koeficientus, žymėjimą ar / ir funkcijas, grafikus. Iš viso koeficientais skyrėsi 15 uždavinių, žymėjimu – penki, funkcijomis / grafikais – keturi, o du uždaviniai buvo identiški.

Antras testas naudotas 1997 metais Vytauto Didžiojo universitete (VDU) per stojamąjį matematikos egzaminą. Jo buvo keturi variantai, kiekvienas po 16 uždavinių. Pirmieji 12 uždavinių buvo su pasirenkamaisiais atsakymais, keturių reikėjo užrašyti atsakymą bei sprendimą. Skirtingų testo variantų uždavinių su pasirenkamaisiais atsakymais skyrėsi klaidinančių ir teisingų atsakymų išdėstymo tvarka, o dviejų paskutinių uždavinių – ir skaitiniai koeficientai. Statistinei analizei naudosime visų stojančiųjų matematikos egzamino rezultatus (829 stojantieji testo variantus sprendė taip: I variantą – 231, II – 228, III – 138 ir IV – 232).

Trečias šiame straipsnyje tiriamas testas yra iš analogiško 1998 metų VDU stojamojo matematikos egzamino. Testo struktūra nesikeitė – 12 uždavinių su pasirenkamaisiais atsakymais, keturių reikėjo užrašyti atsakymą ir sprendimą. Testo variantai sudaryti taip: pirmų 12 skyrėsi uždavinių išdėstymo tvarka. Kaip ir 1997 metų teste, skirtingų variantų uždavinių su pasirenkamaisiais atsakymais atsakymai buvo išdėstyti kita tvarka. Analizei naudojami visų stojančiųjų matematikos egzamino rezultatai (iš 886 stojančiųjų: I testo variantą sprendė – 295, II variantą – 295 ir III variantą – 296).

Kaip matome, šiame straipsnyje pasirinktų testų variantų sudarymo būdai apima pačius paprasčiausius ir dažniausiai naudojamus – pakeičiama atsakymų tvarka, skaitiniai koeficientai arba uždavinių tvarka. Keliais testų teorijoje naudojamais statistiniais metodais bandysime patikrinti, kokią įtaką šie pakeitimai gali turėti testo variantų ekvivalentumui.

Paprasčiausias skirtingų testo variantų palyginimo būdas yra variantų rezultatų skirstinių ir jų vidurkių palyginimas. Vidurkių skirtumo statistiniam patikimumui tikrinti naudosime Stjudento kriterijų nepriklausomoms imtims. Esant pakankamai didelėms imtims, Stjudento kriterijiu taikyti pakanka sąlygos, kad abi lyginamos imtys būtų atsitiktinai parinktos iš dviejų nepriklausomų grupių. G. W. Bonhrnstedas ir D. Knoke (Bonhrsted G. W., Knoke D., 1994, p. 139) nurodo, kad mažos imtys, kai Stjudento kriterijui taikyti būtina normalaus skirstinio sąlyga, yra tokios, kai jų dydžių suma yra mažesnė už 100. Visais straipsnyje tiriamais atvejais imtys yra pakankamai didelės (daugiau nei 800 respondentų), o atskirus variantus sprendusių respondentų grupės yra nepriklausomos. Testo variantai visais atvejais buvo priskirti respondentams atsitiktinai. Atskirų kintamųjų vidurkių skirtumus laikysime statistiškai reikšmingais, kai Stjudento kriterijaus reikšmingumo lygmuo bus mažesnis už 0,05 (žymėsime $p < 0,05$).

Atvejais, kai uždavinio įvertinimas yra tik „teisingai" arba „neteisingai", lyginsime teisingai atsakiusiųjų respondentų dalis (procentais). Kai uždavinio įvertinimas yra taškais (pvz., 0, 1, 2 arba 3), lyginsime kiekvieno varianto taškų skirstinius. Skirtumų tarp teisingai atsakiusiųjų respondentų dalių (procentais) ir taškų skirstinių statistiniam patikimumui tikrinti taikysime Chi^2 (Chi kvadrato) kriterijų. Prie lentelių Chi^2 statistikos reikšmė bus žymima Chi^2, o skliaustuose nurodytas laisvės laipsnių skaičius. Chi^2 kriterijaus statistinio reikšmingumo lygmuo žymimas p raide. Chi^2 kriterijui naudoti paprastai reikia dviejų sąlygų – didelio stebėjimų skaičiaus ir gana didelės dvimatės lentelės ląstelių tikėtinos reikšmės. Daugelis autorių nurodo, kad pirmai sąlygai patenkinti pakanka stebėjimų skaičiaus didesnio nei 20, o antrai sąlygai – kad ląstelių su tikėtinomis reikšmėmis mažesnėmis už 5, nebūtų daugiau nei 20 proc. Šiame tyrime stebėjimų skaičius gerokai viršija 20, o ląstelių su mažomis reikšmėmis skaičius nė vienu atveju neviršijo 20 proc., todėl minėtų sąlygų tenkinimo kiekvienu atveju nerodysime.

Be vidutinių reikšmių, lyginimo reikia atkreipti dėmesį ir į galimus variantų rezultatų skirtumus visoje tiriamų gebėjimų ar žinių lygio skalėje. Todėl atskirais atvejais naudosimės grafiniu testo variantų rezultatų kvantilių palyginimu.

Šiuolaikinė testų teorija neapsiriboja klasikinių statistikos metodų taikymu. IRT teorijos metodais testuojamųjų gebėjimai yra įvertinami ne tik pagal bendrą teisingai atsakytų klausimų skaičių, bet ir pagal tai, į kokio sunkumo klausimus atsakyta teisingai. Taikant šiuos metodus vertinami klausimų ar užduočių parametrai (sunkumas, skiriamoji geba, teisingo atsakymo atspėjimo šansas ir kiti), kurie po to naudojami vertinant testuojamųjų gebėjimus. Šiame straipsnyje statistinei analizei naudosime dviejų parametrų IRT modelį, kuris remiasi prielaida, kad kiekvienas testo klausimas (uždavinys) gali būti nusakomas dviem parametrais – sunkumu ir skiriamąja geba. Tuomet šio modelio pagrindas yra lygtis, nusakanti tam tikrų gebėjimų testuojamojo tikimybę teisingai išspręsti uždavinį:

$$P(\theta) = \frac{e^{Da(\theta - b)}}{1 + e^{Da(\theta - b)}},$$

čia: θ – gebėjimų įvertis, b – uždavinio (klausimo) sunkumas, a – uždavinio skiriamoji geba (parametras, nusakantis uždavinio galią atskirti gabiuosius nuo negabių testuojamų gebėjimų atžvilgiu) ir D – normuojanti konstanta. Parametrams rasti dažniausiai naudojami artutiniai metodai, kurie įvertina atskirų uždavinių parametrus bei testuojamųjų gebėjimus. Gebėjimų įvertis θ paprastai yra išreiškiamas *z*-taškų skalėje, t. y. vidurkiu, lygiu 0, ir standartiniu kvadratiniu nuokrypiu, lygiu 1. Detaliau apie IRT modelius ir jų taikymą žr. [8].

Duomenų analizei naudosimės SPSS programiniu paketu ir specializuota testų analizės programa XCALIBRE™, kuri dviejų parametrų IRT modelio parametrams rasti naudoja marginalinį didžiausio tikėtinumo metodą. Tyrimo analizei naudosime ne pačius uždavinių parametrus, o informacines funkcijas $I(\theta)$, kurios nesunkiai išvedamos iš pagrindinės modelio lygties. Visų testo uždavinių informacinių funkcijų suma yra vadinama testo informacine funkcija. Atskiro uždavinio informacinė funkcija neturi praktinės interpretacijos. Testo informacinė funkcija yra siejama su testo įverčių tikslumu, t. y. kaip tiksliai testas įvertina skirtingų gebėjimų respondentus. Turėdami testo variantų informacines funkcijas palyginsime jų grafikus, ar vienodai tiksliai lygiagretūs variantai įvertina testuojamuosius (išsamiau apie testo informacines funkcijas ir jų taikymą žr. [8; 13].

Tyrimo rezultatai

2 lentelė. ***Tiriamų testų rezultatai***

		Vidurkis ir standartinis kvadratinis nuokrypis	*Skirtumas tarp variantų vidurkių*		
			I variantas	*II variantas*	*III variantas*
Bandomojo matematikos egzamino testas	*I variantas*	18,4 (10,03)	–	–	–
	II variantas	18,8 (10,03)	0,4 (p>0,05)	–	–
VDU 1997 m.	*I variantas*	12,5 (4,53)	–	–	–
	II variantas	12,5 (5,19)	0,04 (p>0,05)	–	–
	III variantas	13,0 (4,76)	0,5 (p>0,05)	0,5 (p>0,05)	–
	IV variantas	13,3 (4,78)	0,8 (p>0,05)	0,8 (p>0,05)	0,3 (p>0,05)
VDU 1998 m.	*I variantas*	12,4 (5,17)	–	–	–
	II variantas	13,1 (4,92)	0,7 (p>0,05)	–	–
	III variantas	13,4 (5,08)	**1,0 (p<0,05)**	0,3 (p>0,05)	–

Nė vienu nagrinėjamu variantų sudarymo atveju, lyginant testo variantų rezultatų vidurkius, skirtumai tarp jų statistiškai nėra reikšmingi, kai reikšmingumo lygmuo p = 0,05 (2 lentelė). Išimtis yra VDU 1998 metų testo I ir III variantai. Šiame teste jau vieno taško skirtumas yra statistiškai reikšmingas. Nesant skirtumų tarp vidutinių reikšmių, galima prielaidą, kad testų, kurie naudojami tik apibendrintai analizei, gali būti naudojami visi nagrinėti variantų sudarymo būdai, o taip sudarytų variantų rezultatai yra palyginami tarpusavyje be pataisų. Tokiais atvejais naudojamų vidutinių reikšmių skirtumai gali būti atsitiktiniai ir statistiškai nereikšmingi.

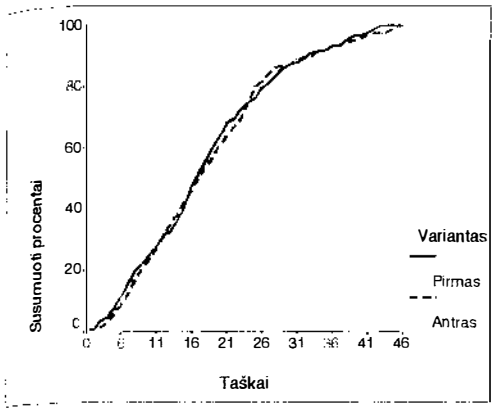
Norint išaiškinti, kodėl skirtumas tarp VDU 1998 m. testo I ir III variantų yra statistiškai reikšmingas, reikia nagrinėti testą sudarančius uždavinius ir naudoti kitokius statistinius, pavyzdžiui, IRT teorijos, metodus.

Kol kas lygindami ieškojome tik vidutinių rezultatų skirtumų. Norint įsitikinti, ar variantai ekvivalentūs visoje gebėjimų skalėje, galima palyginti kvantilius. Diagramose (1–3 diagramos) horizontalioje x ašyje atidėti atitinkamų testo variantų rezultatai (testo taškais), o vertikalioje y ašyje kiekvieną taškų skaičių atitinkantis procentinis rangas, t. y. kiek procentų testo variantą atlikusių moksleivių gavo nedidesnį už tą taškų skaičių įvertinimą.

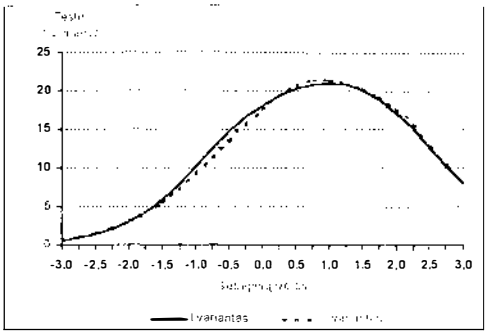
Iš grafikų galima pastebėti, kad skirtingus testo variantus sprendę ir vienodai taškų surinkę gauna ne tą patį procentinį rangą, t. y. jų vieta rezultatus rikiuojant pagal testo taškus skiriasi.

Didesnių bandomojo matematikos testo variantų skirtumų nėra. Labai nedideli skirtumai (nutolusios viena nuo kitos kreivės) matomi beveik visoje testo taškų skalėje abiejų stojamųjų testų – VDU 1997 metais ir VDU 1998 metais. Tokius rangų neatitikimus gali lemti minimalūs testo variantų uždavinių skirtumai. Todėl, jei galutiniai įvertinimai yra tiesiogiai susiję su rangais, tokie skirtumai daro testo rezultatus netikslius ir mažiau patikimus.

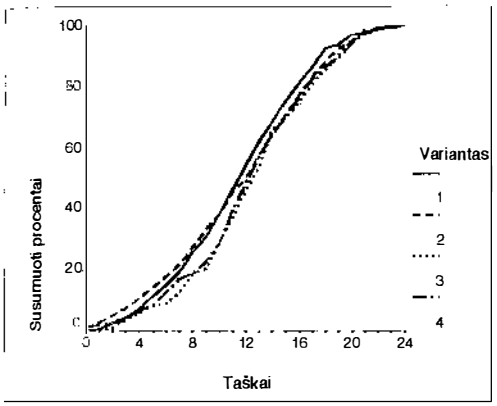
Dar vienas būdas palyginti lygiagrečius variantus – testo informacinės kreivės žr. [13]. Jas lyginant galima įvertinti, kaip tiksliai įvertinami



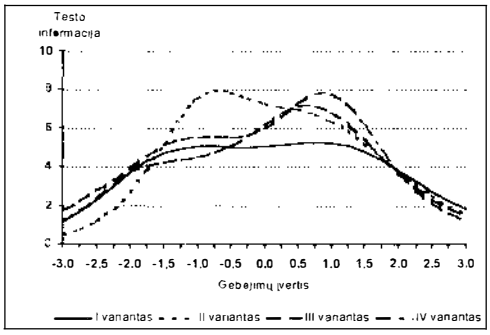
1 diagrama. ***Bandomojo matematikos egzamino testo rezultatai***



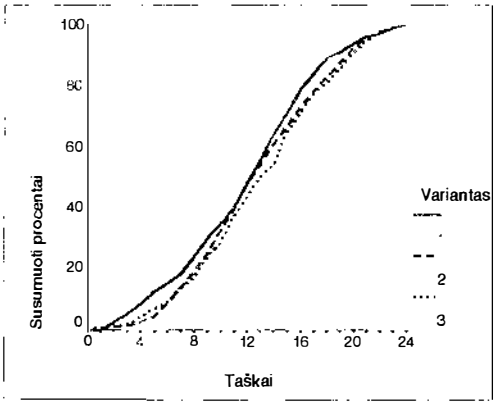
4 diagrama. ***Matematikos bandomojo egzamino testo variantų informacinės kreivės***



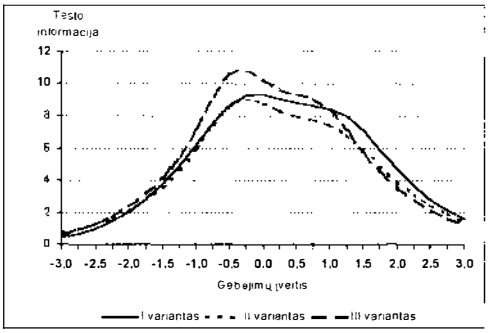
2 diagrama. ***VDU 1997 m.***



5 diagrama. ***Stojamojo VDU 1997 m. matematikos egzamino testo variantų informacinės kreivės***



3 diagrama. ***VDU 1998 m.***



6 diagrama. ***Stojamojo VDU 1998 m. matematikos egzamino testo variantų informacinės kreivės***

tikrieji testuojamųjų gebėjimai visoje skalėje. Tačiau ir šiuo atveju bandomojo matematikos testo ir stojamojo VDU 1998 metų testo kreivės ypač nesiskiria, o VDU 1997 metų testo variantai skiriasi gerokai labiau.

Analizuodami testo variantų informacines kreives matome, kad kai klasikinės testų teorijos metodai neparodo testų variantų skirtumų, moderniosios testų teorijos analizės metodai gali juos parodyti. Tai akivaizdžiai parodo VDU 1997 metų stojamasis testas, kuriame išsiskiria visų variantų informacijos kiekis gana plačioje vidurinėje gebėjimų skalės juostoje.

Žinoma, kad vieno uždavinio informacijos kiekio maksimumas yra pasiekiamas gebėjimų skalės taške, atitinkančiame to uždavinio sunkumą, ir yra tiesiogiai proporcingas uždavinio skiriamosios gebos (modelio funkcijos parametro a) kvadratui (Hambelton R. K., Swaminathan H., 1990, p. 105). Kadangi testo informacijos funkcija yra atskirų uždavinių informacijos funkcijų suma, tai atsiradusius skirtingų testo variantų skirtumus galima paaiškinti uždavinių sunkumo (modelio funkcijos parametras b) ir skiriamosios gebos parametrų skirtumais, t. y. kad uždavinių, kurių aukšta skiriamoji geba, sunkumas skirtingų testo variantų skiriasi. VDU 1997 m. testo variantai buvo sudaromi sukeičiant vietomis uždavinių su pasirenkamaisiais atsakymais atsakymų alternatyvas. Taigi net toks paprastas uždavinio formuluotės pakeitimas gali lemti gana didelius uždavinio statistinių parametrų skirtumus.

Kiekvieno testuojamojo gebėjimų ar žinių įverčiams turi įtakos ne tik bendras pasiektas rezultatas (t. y. teisingai išspręstų uždavinių skaičius), bet ir informacija, kuriuos (kokio sunkumo ar kitų statistinių charakteristikų) uždavinius jis išsprendė teisingai. Testo variantų statistinių charakteristikų skirtumai, pastebėti IRT teorijos metodais, verčia variantus analizuoti detaliau, palyginti atskirus uždavinius. Šiame straipsnyje nagrinėjamus testus galima palyginti, nes testo variantai sudaryti taip, kad kiekvienas varianto uždavinys turi kito varianto atitikmenį. Šie du uždaviniai yra traktuojami kaip ekvivalentūs ir *a priori* turėtų turėti tas pačias statistines charakteristikas.

Palyginus visų nagrinėjamų testų tam tikrų variantų užduotis, išsiskiria aštuoni uždaviniai, kurių parametrų skirtumai yra statistiškai reikšmingi. Kiti testo variantų uždaviniai statistiškai nesiskiria. Tad panagrinėkime tik šiuos išsiskiriančius uždavinius, palyginkime statistines jų charakteristikas.

3–10 lentelėse šie uždaviniai pateikiami greta. Jei skiriasi tik alternatyvių atsakymų tvarka, uždavinio sąlyga abiem variantams yra pateikiama viena. Pateikiama uždavinius teisingai išsprendusiųjų dalis (procentais) ir statistinis variantų parametrų skirtumo įvertis – Chi^2 statistika su reikšmingumo lygmeniu. Alternatyvius atsakymus pasirinkusiųjų dalis (procentais) yra pateikiama po kiekvieno atsakymo, o teisingas atsakymas pažymėtas žvaigždute ir išskirtas juodesniu šriftu.

3 lentelė. ***Matematikos bandomojo egzamino testo 9a uždavinys***

I variantas	*II variantas*
Paleistas iš 6,25 m aukščio rutuliukas kaskart atšokdamas nuo pagrindo netenka 40% prieš tai buvusio aukščio. Į kokį aukštį rutuliukas pakils atšokęs nuo pagrindo pirmąjį kartą?	Paleistas iš 2,56 m aukščio rutuliukas kaskart atšokdamas nuo pagrindo netenka 25% prieš tai buvusio aukščio. Į kokį aukštį kamuoliukas pakils atšokęs nuo pagrindo pirmąjį kartą?
Teisingai išsprendė **70,3%** moksleivių.	Teisingai išsprendė **81,9%** moksleivių.

$Chi^2 = 7{,}973$ (1), $p = 0{,}005$

Šis uždavinys testo variantuose skiriasi tik skaitiniais koeficientais. Po tokio pakeitimo I variantą sprendę moksleiviai padarė daugiau skaičiavimo klaidų, jiems skaičiavimas galėjo būti mažiau „patogesnis“: 25%, t. y. skaičiaus ketvirtis skaičiuojamas lengviau nei 40%. Šis pavyzdys rodo, kad net ir iš pirmo žvilgsnio mažai reikšmingas skaičiaus pakeitimas, nekeičiant viso uždavinio sąlygos, gali padaryti uždavinį sunkesnį, ir moksleivio pažymiui didesnę reikšmę turės atsitiktinumas, kuris variantas jam teks per egzaminą.

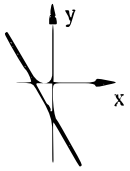
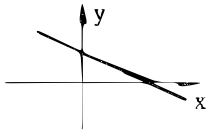
4 lentelė. ***Matematikos bandomojo egzamino testo 12 uždavinys***

I variantas			*II variantas*		
$\sin 3x \sin 2x - \cos 3x \cos 2x =$			$\cos 4x \sin x - \sin 4x \cos x =$		
A	$\sin x$	(10,1%)	A	$\cos 5x$	(4,2%)
B	$\cos(-x)$	(8.4%)	B	$\sin 5x$	(5.9%)
C	$-\cos x$	(15,2%)	C	$\sin 3x$	(26,5%)
D	$\cos 5x$	(23,6%)	D	$-\sin 5x$	(4,6%)
E*	$\mathbf{-\cos 5x}$	**(42,6%)**	**E***	$\mathbf{-\sin 3x}$	**(58,8%)**

$Chi^2 = 11{,}801$ (1), $p = 0{,}001$

Šiame uždavinyje skiriasi trigonometrinės funkcijos, tačiau abiem atvejais sprendžiant uždavinį reikia pasinaudoti labai panašiomis trigonometrinėmis tapatybėmis. Visos reikalingos formulės pateikiamos testo sąsiuvinyje. Pirmo varianto neteisingi atsakymai buvo pasirinkti dažniau nei antro, todėl teisingai uždavinį išsprendė mažiau moksleivių nei sprendę antrą variantą. Galima daryti prielaidą, kad šį variantų skirtumą nulėmė tai, kad sprendimams reikėjo taikyti skirtingas tapatybes ir I varianto sprendimas buvo sudėtingesnis.

5 lentelė. ***Matematikos bandomojo egzamino testo 13 uždavinys***

I variantas			*II variantas*		
Kurios funkcijos grafiko eskizas pavaizduotas paveiksle? 			Kurios funkcijos grafiko eskizas pavaizduotas paveiksle? 		
A*	$y = -2x - 3$	**(71,8%)**	A	$y = -\frac{1}{2}x - 1$	(8,3%)
B	$y = -2x + 3$	(8,8%)	**B***	$y = -\frac{1}{2}x + 1$	**(52,5%)**
C	$y = 2x + 3$	(9,2%)	C	$y = \frac{1}{2}x + 1$	(28,5%)
D	$y = -2x$	(2,5%)	D	$y = -\frac{1}{2}x$	(1,7%)
E	$y = 2x -$	(7,6%)	E	$y = \frac{1}{2}x - 1$	(9,1%)

$Chi^2 = 17{,}510$ (1), $p < 0{,}001$

Šiame uždavinyje buvo pakeistas grafiko eskizas. Uždavinio sprendimui toks pakeitimas įtakos neturi – abiem atvejais reikia tų pačių žinių ar gebėjimų bei taikyti tuos pačius veiksmus. II varianto C atsakymą pasirinko 28,5% moksleivių. I variante tokio atsakymo nėra. Taigi, nesigilinant į matematines uždavinio turinio ypatybes, kaip vieną iš realiausių variantų skirtumo paaiškinimų būtų galima nurodyti statistiškai nelygiaverčių atsakymų parinkimą. Tokio efekto galima išvengti, jei testas, prieš pateikiant moksleiviams, būtų bandomas ir klaidinantys atsakymai galėtų būti derinami atsižvelgiant į statistines jų charakteristikas.

6 lentelė. ***VDU 1997 m. stojamojo matematikos testo 3 uždavinys***

I variantas		*II variantas*		*III variantas*		*IV variantas*	
Lygties $\sqrt{5x} - \sqrt{3x} = 2$ sprendinys yra							
A) 2	(13,0%)	A) 2	(10,1%)	A) 2	(8,1%)	A) 2	(5,9%)
B) $\sqrt{2}$	(2,2%)	B) $\sqrt{2}$	(1,7%)	**B) $8+2\sqrt{15}$**	**(69,1%)**	B) $\sqrt{2}$	(1,1%)
C) $8-2\sqrt{15}$	(25,0%)	C) $\sqrt{5}+\sqrt{3}$	(6,2%)	C) $8-2\sqrt{15}$	(19,1%)	**C) $8+2\sqrt{15}$**	**(72,7%)**
D) $8+2\sqrt{15}$	**(57,6%)**	D) $8-2\sqrt{15}$	(11,2%)	D) $\sqrt{2}$	(2,2%)	D) $8-2\sqrt{15}$	(15,0%)
E) $\sqrt{5}+\sqrt{3}$	(2,2%)	**E) $8+2\sqrt{15}$**	**(70,8%)**	E) $\sqrt{5}+\sqrt{3}$	(1,5%)	E) $\sqrt{5}+\sqrt{3}$	(5,3%)

$Chi^2 = 11{,}676$ (3), $p = 0{,}009$

VDU 1997 m. testo variantai buvo sudaryti iš tų pačių uždavinių, išdėstytų tose pačiose testo vietose, o uždavinių su pasirenkamais atsakymais atsakymai (ir teisingasis, ir klaidinantys) sukeisti vietomis. Šiame uždavinyje matome, kaip keitėsi atsakymų vieta. Didžiausi pastebimi pirmojo varianto ir likusių skirtumai. II, III ir IV variantų skirtumai nėra dideli. Šiame pavyzdyje matome, kad atsakymą pasirinkusių moksleivių dalis skiriasi ne tik dėl to atsakymo vietos tarp visų atsakymų (pvz., A atsakymas). Tai rodo, kad atskirų testo variantų uždavinio sunkumo skirtumui įtakos turi ne tik vieno atsakymo vieta, bet ir visų jų išdėstymo tvarka.

7 lentelė. ***VDU 1997 m. stojamojo matematikos testo 12 uždavinys***

I variantas		*II variantas*		*III variantas*		*IV variantas*	
Didžiausioji funkcijos $y = 2x^3 + 3x^2 - 72x$ reikšmė atkarpoje $[-6;6]$ yra							
A) 108	(35,6%)	A) 108	(36,7%)	A) 108	(36,2%)	A) 108	(28,1%)
B) 208	**(49,2%)**	B) 720	(3,7%)	B) 497	(3,7%)	B) 185	(3,6%)
C) 185	(9,9%)	C) 185	(5,9%)	C) 185	(5,9%)	**C) 208**	**(64,1%)**
D) 497	(1,1%)	D) 497	(3,7%)	**D) 208**	**(52,2%)**	D) 497	(2,6%)
E) 720	(4,2%)	**E) 208**	**(50,0%)**	E) 720	(1,4%)	E) 720	(1,6%)

$Chi^2 = 11{,}524$ (3), $p = 0{,}009$

Šis pavyzdys panašus į jau nagrinėtą 4, tik labiausiai išsiskiria ne pirmas, o ketvirtas variantas. Testo variantų rezultatų skirtumas taip pat galėtų būti aiškinamas skirtingo alternatyvių atsakymų derinio poveikiu.

VDU 1998 m. stojamojo matematikos testo variantai buvo sudaryti ne tik sumaišant alternatyvius atsakymus, bet ir sukeičiant vietomis pačius uždavinius.

8 lentelė. ***VDU 1998 m. stojamojo matematikos testo 1 uždavinys***

I variantas		*II variantas*		*III variantas*	
Žinoma, kad reiškinys $\sqrt{11+6\sqrt{2}}+\sqrt{11-6\sqrt{2}}$ lygus sveikajam skaičiui. Kuriam?					
A 9	(4,8%)	A 9	(2,7%)	A 9	(1,7%)
B 4	(4,1%)	B 11	(10,9%)	B 11	(7,1%)
C 6	**(71,0%)**	**C 6**	**(76,9%)**	**C 6**	**(80,6%)**
D 5	(6,5%)	D 5	(6,8%)	D 5	(6,5%)
E 11	(13,7%)	E 4	(2,7%)	E 4	(4,1%)
Uždavinio vieta teste:					
1		*3*		*6*	

$Chi^2 = 7{,}531$ (2), $p = 0{,}023$

Šį uždavinį blogiausiai sprendė I variantą gavę egzaminuojamieji, t. y. kai uždavinys teste buvo pirmasis. Teisingas atsakymas buvo toje pačioje vietoje, tačiau skiriasi I ir kitų variantų klaidinančių atsakymų išdėstymo tvarka. Šiek tiek santykiškai daugiau pasirinkti I varianto pirmasis ir paskutinysis (A ir E) atsakymai ir nulėmė teisingo atsakymo pasirinkimo skirtumus.

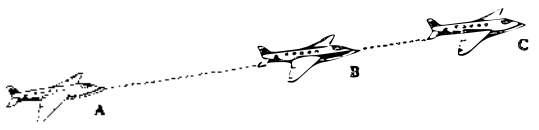
Kitų dviejų uždavinių skyrėsi ne tik išdėstymo vieta teste, bet ir alternatyvių atsakymų išdėstymo tvarka. Sudėtinga įvertinti, kiek įtakos variantų rezultatų skirtumui turi uždavinio vieta teste ir kiek atsakymų išdėstymas, tačiau net ir esant tik tokiems pakeitimams atsirado statistiškai reikšmingi teisingai išsprendusių atskirų variantų uždavinius respondentų dalies skirtumai.

9 lentelė. ***VDU 1998 m. stojamojo matematikos testo 3 uždavinys***

I variantas		*II variantas*		*III variantas*	
Kiek šaknų atkarpoje $[0;\ 2\pi]$ turi lygtis $\sin x + \cos x = 1$?					
A 6	(3,4%)	A 6	(3,7%)	A 6	(3,1%)
B 5	(15,9%)	B 5	(26,1%)	B 4	(13,6%)
C 2	(36,6%)	C 2	(30,5%)	**C 3**	**(38,6%)**
D 3	**(28,1%)**	**D 3**	**(28,8%)**	D 2	(29,2%)
E 4	(15,9%)	E 4	(10,8%)	E 5	(15,6%)
Uždavinio vieta teste:					
3		*2*		*3*	

$Chi^2 = 9{,}186$ (2), $p = 0{,}010$

10 lentelė. ***VDU 1998 m. stojamojo matematikos testo 8 uždavinys***

I variantas		*II variantas*		*III variantas*	
Lėktuvas skrisdamas pastoviu greičiu (ir pastovia kryptimi) iš taško A į tašką B skrenda 2 min., iš taško B į tašką C – 1 min. Įvedus koordinačių sistemą, taškų A ir B koordinatės bus atitinkamai (–1; 4; 1) ir (3; –2; 3). Taško C koordinatės bus: 					
A (5; –5; 4)	**(49,0%)**	**A (5; –5; 4)**	**(58,5%)**	A (4; –6; 2)	(15,6%)
B (7; –8; 5)	(11,9%)	B (4; –6; 2)	(9,2%)	B (7; –8; 5)	(6,8%)
C (2; 2; 4)	(16,0%)	C (2; 2; 4)	(15,6%)	C (2; 2; 4)	(12,9%)
D (4; –6; 2)	(16,7%)	D (7; –8; 5)	(8,5%)	**D (5; –5; 4)**	**(60,0%)**
E (5; –3; 5)	(6,5%)	E (5; –3; 5)	(8,2%)	E (5; 3; 5)	(4,7%)
Uždavinio vieta teste:					
8		*7*		*8*	

$Chi^2 = 8{,}491$ (2), $p = 0{,}014$

Išvados

Straipsnyje nagrinėti tik paprasčiausi testų variantų sudarymo būdai – tiktai minimalūs uždavinio formos pakeitimai. Taip gaunami testo variantai yra ekvivalentaus turinio ir turi panašias bendrąsias statistines charakteristikas: vidurkius, dispersijas ir pan. Tačiau nagrinėjant detaliau, t. y. uždavinių statistinių charakteristikų skirtumus bei taikant moderniosios testų teorijos IRT analizės metodus, variantų ekvivalentumas net ir tokiomis situacijomis tampa abejotinu. Ir minimalūs uždavinio sąlygos pakeitimai gali diskriminuoti juos sprendžiančius moksleivius. Šiuo požiūriu visiškai ekvivalenčių vieno uždavinio netrivialių variantų nėra.

Atsirandančių testo variantų skirtumų įtaką galima sušvelninti bandomojo testavimo metu kruopščiau parenkant kitą uždavinio variantą arba jau po testavimo naudojant specialius išlyginamuosius statistinius metodus (pvz., Rasho skales, kvantilių sulyginimą ar kt.). Bandomasis testavimas dažnai neatliekamas dėl didelių išlaidų arba siekiant išvengti išankstinio informacijos paskleidimo. Išlyginamieji statistiniai metodai Lietuvoje dar nėra taikomi. Šiems metodams reikia ne tik tinkamo testų rengėjų profesinio pasirengimo, bet ir taikyti specialią (taip pat nemažai kainuojančią) programinę įrangą.

Nesant didesnių vidurkių ir dispersijų skirtumų, galima teigti, kad minėtais būdais sudarytus variantus yra tinkama naudoti diagnostiniam testavimui, vertinant bendras atskirų grupių žinių ar gebėjimų charakteristikas.

Tačiau kai testo rezultatas yra daug lemiantis testuojamajam, reikia vengti naudoti skirtingus testo variantus, o testo atlikimo sąžiningumą stengtis užtikrinti kitais būdais. Per egzaminą dažnai vienintelis taškas gali nulemti testuojamojo vietą ir, jei pagrindinis veiksnys, nulėmęs to taško praradimą, yra variantų skirtumas, moksleivis tikrai gali kaltinti egzamino rengėjus, kad buvo gavęs „sunkesnį" testo variantą.

Už vertingas pastabas rengiant šį straipsnį dėkoju mokslinio darbo vadovui docentui Algirdui Zabulioniui.

LITERATŪRA

1. Angoff W. A. Scales, norms, and equivalent scores // Thorndike R. L. (ed.) Educational Measurement, 2nd edn. American Council on Education, 1971.

2. Bohrnstedt G. W., Knoke D. Statistics for Social Data Analysis. 3rd ed. F. E. Peacock Publishers, Inc., 1982.

3. Brennan R. L. (ed.) Problems, perspectives and practical issues in equating // Applied Psychological Measurement, 1987, 11(3).

4. Brennan R. L., Kolen M. J. Some Practical Issues in Equating // Applied Psychological Measurement, 1987, 11(3).

5. Cizek G. J. Cheating On Tests. How To Do It, Detect It, And Prevent It. Lawrence Earlbaum Associates, 1999.

6. Cook L. L., Eignor D. R. NCME instructional module: IRT equating methods. Educational Measurement: Issues and Practice, 1991, 10(3).

7. Čekanavičius V., Trakas G., Zabulionis A. Trečioji tarptautinė matematikos ir gamtos mokslų studija. 7–8 klasių moksleivių tyrimo statistinė ataskaita. Vilnius, 1997.

8. Hambelton R. K., Swaminathan H. Item Response Theory. Principles and Applications. Kluwer-Nijhoff Publishing, 1990.

9. Holland P. W., Rubin D. B. (eds.) Test Equating. Academic Press, 1982.

10. Hollinger R. C. & Lanza-Kaduce L. Academic Dishonesty and the Perceived Effectiveness of Countermeasures: An Empirical Survey of Cheating at Major Public University // NASPA Journal, 1996, 33(4).

11. Kolen M. J. An NCME instructional module on traditional equating methodology. Educational Measurement: Issues and Practice, 1988, 7(4).

12. Kolen M. J., Brennan R. L. Test Equating. Methods and Practices. Springer-Verlag, 1995.

13. Trakas G. Testo informacija ir jos taikymai // Informacijos mokslai. Vilniaus universitetas, 1997. Nr. 7.

PROBLEMS OF THE EQUIVALENCE OF THE PARALLEL TEST FORMS

Gediminas Trakas

Summary

The importance of correct usage of the parallel test forms is evident not only in the field of educational diagnostic testing. Especially it applies in „high stakes" examinations where the important efforts for honesty should not violate the principles of equal opportunities of students. Some simple methods of the development of parallel test forms are investigated in this article. The results showed, that in some cases simple changes to the item format (e.g. changing the order of distracters, changing coefficients and similar) causes statistically significant differences in parameters between the parallel items. Such differences have no effect on the average scores, however they may become significant in other parts of the ability scale. Some applications of the Item Response Theory (IRT) methods have been used to compare parallel test forms.

Gauta 2000 10 17
Priimta 2000 11 23