СЪПОСТАВИТЕЛНО ИЗСЛЕДВАНЕ НА ОЧАКВАНОТО „ПОВЕДЕНИЕ” НА СТАТИСТИКИТЕ НА ТЕСТОВИТЕ ВЪПРОСИ, ОПРЕДЕЛЕНИ В РАМКИТЕ НА КЛАСИЧЕСКАТА ТЕСТОВА ТЕОРИЯ И ТЕОРИЯТА ЗА ОТГОВОР НА ТЕСТОВ ВЪПРОС

Л. Джалев

Нов български университет

Резюме

Класическата тестова теория и Теорията за отговор на тестов въпрос са алтернативни, конкурентни системи за оценяване на латентни черти. Въпреки съществените различия между тях, те споделят някои общи теоретични конструкти, каквито са характеристиките на въпросите трудност, дискриминативна сила и налучкване на правилния отговор. Всяка теория съществува под формата на множество модели, които могат да бъдат разглеждани като модели на данни. От своя страна, моделите следва да бъдат оценени от гледна точка на тяхната приложимост към конкретна съвкупност от тестови данни. Един от подходите за извършване на такава оценка е да се провери дали особеностите на съответния модел ще се проявят, когато бъде приложен към тези данни.

Новата психометрична теория има редица теоретични предимства пред Класическата тестова теория. По отношение на параметрите на тестовите въпроси те се изразяват в прецизността на техните оценки, независимостта им от извадката, независимостта им една от друга. Тези характеристики, но в противоположен смисъл, се разглеждат като съществени недостатъци на Класическата теория.

Обект на настоящото изследване е очакваното „поведение” на статистиките на въпросите, определени в рамките на два теоретични модела, приложени върху данни от тестове за постижения. Формулирани са три групи от допускания в съгласие с теоретичните очаквания - за инвариантността на статистиките, за взаимовръзките между статистиките в рамките на един и същи модел, както и за съгласуваността между съответстващите си статистики от двата модела.

Макар че психометричната общност взема предимствата на Теорията за отговор на тестов въпрос за даденост, резултатите от направените изследвания поставят по съмнение повечето от тях. Параметрите на въпросите не са инвариантни в очакваната степен, между

тях съществуват силни взаимовръзки, а между едноименните статистики от двата модела се наблюдава висока степен на съгласуваност. И обратно – индексите в рамките на Класическата тестова теория демонстрират поведение, което е съпоставимо с това на параметрите, а в някои отношения дори ги превъзхождат.

Ключови думи: *Класическата тестова теория, Теория за отговор на тестов въпрос, индекси, параметри, трудност, дискриминативна сила, налучкване на правилния отговор, коефициент на корелация, корелационно отношение, диаграма на разсейване*

I. Постановка на емпиричното изследване

1. Цели на изследването

Изследването на очакваното „поведение" на статистиките на тестовите въпроси, определени в рамките на двете основни психометрични теории, е част от едно по-широко съпоставително изследване на тяхната приложимост към резултатите от Теста по общообразователна подготовка (ТОП), който стои в основата на приемните процедури в Нов български университет.

Класическата тестова теория (CTT) и Теорията за отговор на тестов въпрос (IRT) могат да бъдат разглеждане като конкурентни теории, защото (а) имат едно и също поле на приложение, (б) са средства за изследване на едни и същи латентни черти, (в) моделите, разработени в техните рамки, са модели на едни и същи данни и (г) са функционално еквивалентни, но независими, алтернативни една на друга. Оттук произтича един важен методологически въпрос – дали и в каква степен всяка от тези теории може да бъде използвана като теоретична рамка за поддържане на жизнеността на тази тестова програма.

Един от основните подходи за изследване на приложимостта на дадена тестова теория се основава на оценка на валидността на нейните основни допускания по отношение на тестовите данни. Този подход се основава на схващането, че моделите, изградени в рамките на дадена тестова теория, могат да бъдат класифицирани като „модели на данни" (Suppes, 1962; Frigg & Hartmann, 2006). Описвайки този клас модели, П. Супес визира суровите данни, които изследователят получава като непосредствен резултат от проведените от него (емпирични) наблюдения. В този клас модели суровите, реални данни се представят в един непълен, но добре подреден, организиран и в известен смисъл идеализиран вид (Hambleton and Jones, 1993; Frigg and Hartmann, 2006). Нещо повече, „…моделите винаги предлагат непълна репрезентация на тестовите данни, за които са предназначени; по този начин, с достатъчно количество тестови данни, те могат […] да изглеждат непригодни" (Hambleton

and Jones, 1993, стр. 254).

Всеки психометричен модел на данни се базира на определен набор от допускания, които по същество представляват описание на данните, за които е предназначен. Поради тези особености един от фундаменталните въпроси при прилагането на моделите на данни е за връзката, за съответствието, за „съвместимостта" между даден модел, по-точно между неговите допускания, и емпиричните данни, който може да бъде формулиран като въпрос за адекватността на модела (*model-data fit*). В този смисъл един теоретичен модел (в частност моделите на CTT и IRT), следва да се разглежда като приложим, ако има съответствие между допусканията на този модел и съответните характеристика на емпиричните данни. Р. Хамбълтън и Р. Джоунс, разглеждайки проблема за приложимостта на тестовите модели във връзка с техните несъвършенства, отбелязват, че „правилният" въпрос е не дали един модел е правилен или неправилен, а „…дали даден модел съответства на данните достатъчно добре, за да бъде полезен при провеждането на измервателния процес" (Hambleton and Jones, 1993, стр. 254).

Друг подход за изследване на приложимостта на дадена тестова теория, върху който е фокусирано настоящото изследване, е да се направи анализ на очакваните характеристики на даден модел, изграден в нейните рамки. Под „очаквани" характеристики ще разбираме такива особености на модела, които произтичат от начина, по който в него са конкретизирани и детайлизирани основните теоретични конструкти, мрежата от техните взаимовръзки и основни допускания.

Съобразно предназначението, технологията на конструиране, структурата и на начина на използване на резултатите от ТОП, в изследването ще бъдат съпоставени следните два модела: „стандартен" паралелен, τ-еквивалентен модел в рамките на CTT, и едномерен, трипараметричен, логистичен, основан на дихотомични отговори и на нормално разпределение на латентната променлива в рамките на IRT.

Както беше отбелязано, между двете психометрични теории могат да бъдат прокарани множество паралели, един от които свързва статистиките на тестовите въпроси. *CTT* и *IRT* си съперничат и в този аспект, приписвайки един и същи набор от характеристики на тестовите въпроси – трудност,

дискриминативна сила и налучкване на правилния отговор. Разбира се, двете теории не са „изоморфни” по отношение на начина, по който описват тестовите въпроси. Те се различават по отношение на броя на дескрипторите (списъкът с характеристики на въпросите при новата теория е по-дълъг и включва още два параметъра), има разлики между съдържанията на съответните понятия (при новата теория те са по-сложни и формализирани), различават се и по математическите подходи за тяхното оценяване. Като цяло обаче посочените три характеристики имат сходни интерпретации, съответно функции в процеса на конструиране на психометричните инструменти, за анализ и интерпретация на техните резултати.

Тук трябва да уточним, че в това изследване ще се занимаем преди всичко с това какво е очакваното „поведение” на посочените тестови статистики, когато въпросите са поставени в различни условия, и дали наблюдаваното поведение на тези статистики се отклонява от предвиденото. Тези очаквания се основават на дефинициите на съответните характеристики на въпросите в рамките на съответния теоретичен модел, както и на начините за изчисляване на техните стойности. Ще бъдат разгледани и въпросите, свързани със съвместното вариране на едноименните и на разноименните статистики на въпросите, поставени в едно и също условие.

Класическата тестова теория предлага такива методи за оценка на индексите на трудност и дискриминативна сила на въпросите, които биха могли да доведат до зависимост на получените стойности от извадката от изпитани, въз основа на която са изчислени. Такъв тип зависимост е характерна и за резултатите от теста (наблюдавания тестов бал), чийто стойности са подвластни на инструмента за измерване, чрез който са получени. С други думи, може да се очаква, че индексите на въпросите в рамките на *CTT* са нестабилни и вариативни при многократно оценяване върху различни извадки от и. л. Обратно, методите за оценка на параметрите на въпросите в рамките на Теорията за отговор на тестов въпрос предполагат независимостта на параметрите от конкретната извадка от изпитани, въз основа на която са оценени. Може да се очаква, следователно, че оценките на параметрите са стабилни, инвариантни при многократно оценяване върху различни извадки от и. л. Оценката на личностовия параметър Θ също е независима от

инструмента, чрез които е получена.

Например ако с даден тест бъде изпитана група от лица, които имат високи равнища на дадена способност, може да се очаква, че трудността на въпросите в този тест, определена по *CTT*, като цяло ще бъде сравнително ниска. Това предположение се основава на дефиницията, съответно на начина на изчисляване на стойностите на този индекс – като отношение между броя на лицата, отговорили правилно на даден въпрос, и общия брой на лицата, отговорили на този въпрос. Поради това може да се предположи с основание, че делът на правилните отговори ще бъде висок, т. е. трудността на въпросите ще бъде ниска. Ако със същия тест бъдат изпитани лица с ниски равнища на способности, трудността на въпросите като цяло ще бъде висока.

Поради начина на нейното изчисляване в рамките на *CTT*, дискриминативната сила на въпросите също е зависима от извадката от и. л. Конкретните й стойности се определят от дяловете на лицата в силната и в слабата група, отговорили правилно на съответния въпрос. Тези дялове биха могли да се варират, при това значително, при различни извадки от изпитани с различни равнища на способности, особено при нехомогенни извадки.

Такива флуктуации на тестовите статистики биха били невъзможни, ако са определени по методите на *IRT*, т. е. статистиките на въпросите са инвариантни по отношение на извадката от и. л., по-точно от разпределението на способностите Θ в нея (Hambleton, Swaminathan & Rogers, 1991; Fan, 1998; Baker, 2001; Rasch, 2001). Тази особеност произтича от вероятностния подход за изчисляване на стойностите на параметрите и процедурата на максималното правдоподобие за тяхното изчисляване. Авторът на популярния еднопараметричен „Раш" модел Г. Раш отбелязва, че разпределенията на параметрите в този модел са независими и поради това оценката на трудността на въпросите няма да бъде повлияна от това какви стойности на Θ имат индивидите (Rasch, 2001).

Да разгледаме следния пример. Ако характеристичната крива на даден въпрос бъде построена въз основа на извадка, характеризираща се с ниски когнитивни способности, изграждането на тази крива ще бъде направено само за онази част от нея, която обхваща отговорите на лицата от тази група, т. е. частта от характеристичната крива над левия край на скалата Θ. Ако от същата

популация бъде формирана друга извадка с високи когнитивни способности, изграждането на характеристичната крива ще бъде направено само за онази част от нея, която обхваща лицата от тази група, т. е. частта от характеристичната крива над десния край на скалата. И в двата случая оценките на съответните параметри на кривата ще бъдат равни. Това дава основание да се мисли, че „стойностите параметрите са характеристики на въпроса, а не на групата, отговорила на въпроса" (Baker, 2001, стр. 55). Авторите, включително и Ф. Бейкър, все пак оставят вратата отворена, отбелязвайки, че въпреки че действителните стойности на параметрите са инвариантни, техните оценки могат да варират в различните извадки, но в много тесни граници, оставайки почти равни.

Тези особености на статистиките на въпросите се разглеждат като значителен недостатък на Класическата и важно предимство на Теорията за отговор на тестов въпрос. Те обикновено служат и за основания при избора на теоретичен модел за решаване на различни научни или приложни задачи.

Ето защо основната цел на това емпирично изследване е да се потърсят свидетелства за устойчивостта на „поведението" на статистиките на въпросите от Теста по общообразователна подготовка. С други думи, да се установи дали и в каква степен статистиките на въпросите са устойчиви, инвариантни по отношение на извадките от и. л., въз основа на които са получени, или варират при оценка в условията на различни извадки. Поради очакваните различия в поведението на статистиките, обусловено от принадлежността им към една или друга теоретична рамка, тази основна цел ще бъде разложена на три подцели:

1. Да се изследва стабилността (инвариантността) на всяка статистика, определена в рамките на *CTT* и *IRT*, в различни условия, т. е. при различни извадки от и. л.

2. Да се изследват (а) взаимовръзките между индексите на въпросите, определени в рамките на *CTT*, и (б) взаимовръзките между параметрите, определена в рамките на *IRT*, в едно и също условие, т.е. при една и съща извадка от и. л.

3. Да се изследва съгласуваността между индексите, определени в рамките на *CTT*, и съответните им параметри, определена в рамките на *IRT*, в едно и също условие, т.е. при една и съща извадка от и. л.

В анализите няма да бъде включен индексът на налучкване на правилния отговор по *CTT*. Обикновено той се прилага към въпроси с множествен избор и се изчислява като реципрочна стойност на броя на дистракторите. В Теста по общообразователна подготовка всички въпроси са от този вид, с еднакъв брой на дистракторите, поради което при всички въпроси този индекс е с константна стойност, равна на 0.20.

2. Основни допускания

В съгласие с определените цели на изследването и очертаните по-горе очаквания за степента, в която различните статистиките на въпросите, определени съгласно двата теоретични модела, са податливи на изменения в зависимост от извадката, ще направим следните основни допускания:

(1) По отношение на стабилността (инвариантността) на тестовите статистики

(а) Допускаме, че стойностите на индексите на трудност (p) и на дискриминативна сила (D, r_{bis}), определени в рамките на *CTT*, са зависими от извадките, въз основа на които са получени, и поради това тези индекси се характеризират с нестабилност и вариативност. Нестабилността на индексите беше обоснована по-горе в текста.

(б) Допускаме, че стойностите на параметрите на дискриминативна сила (a), трудност (b) и налучкване на правилния отговор (c), определени в рамките на IRT, са независими от извадките, въз основа на които са получени, и поради това тези параметри са стабилни и инвариантни. Стабилността на параметрите също бе обоснована по-горе в текста.

(2) По отношение на (а) взаимовръзките между индексите на въпросите, определени в рамките на *CTT*, и (б) взаимовръзките между параметрите, определена в рамките на *IRT*, в едно и също условие, т.е. при една и съща извадка от и. л.

(а) Допускаме, че между стойностите на индексите на трудност (p) и на дискриминативна сила (D и r_{bis}), определени в рамките на *CTT* върху една и съща извадка, съществува взаимовръзка от нелинеен характер. Допускането

се основава на теорията на данните на К. Кумбс, по-конкретно на модела „Данни единичен стимул" (Coombs, 1964). Ако един въпрос има екстремно висока/ ниска трудност, съгласно теоретичния модел той доминира над/ е доминиран от по-голяма част от индивидите. И в двата случая този въпрос би се характеризирал с екстремно ниска дискриминативна сила. Може да се предположи, че максималните си стойности този индекс би получил при въпроси със средна трудност, позиционирани в средата на скалата на съответния признак, подложен на измерване.

(б) Допускаме, че между стойностите на параметрите на дискриминативна сила (a), трудност (b) и налучкване на правилния отговор (c), определени в рамките на *IRT*, не съществуват взаимовръзки от корелационен вид. Това следва от вероятностния подход на тяхното оценяване, което предполага, че дадена характеристична крива може да заеме различна позиция на скалата на способностите Θ и същевременно да има различен (произволен) наклон или долна асимптота, в границите на изменение на съответния параметър.

(3) По отношение на съгласуваността между съответстващите си индекси и параметри, определени в едно и също условие, т. е. при една и съща извадка от и. л.

Допускаме, че между стойностите на оценките на трудността на въпросите p и b, както и между тяхната дискриминативна сила D и a и r_{bis} и a, няма съгласуваност. Това следва от теоретичната вариативност, нестабилност на индексите, определени в рамките на *CTT*, и тяхната инвариантност в рамките на *IRT*. Ако даден индекс варира в допустимите граници на неговото изменение, а съответният параметър е стабилен, не би могло да се очаква да има съгласуване между техните стойности.

3. Задачи на изследването

Във връзка с формулираната по-горе цели и допускания, следва да бъдат изпълнени следните задачи:

(1) Да се идентифицира съвкупност/ съвкупности от тестови въпроси,

които са изпълнявани от различни (поне две) групи от и. л.

(2) Да се идентифицират тестовите варианти, в които са включени тези въпроси.

(3) Да се формират двойки от тестови варианти, в които са включени едни и същи тестови въпроси.

(4) Да се изчислят индексите на трудност (p) и дискриминативна сила (D), определени в рамките на *CTT*, на въпросите от съответната двойка варианти

(5) Да се изчислят параметрите на трудност (b), дискриминативна сила (a) и налучкване на правилния отговор (c), определени в рамките на IRT, на въпросите от съответната двойка варианти.

(6) Да се приложат адекватни статистически методи за анализ на стабилността/ нестабилността на индексите на въпросите по *CTT* и съответните параметри по *IRT*.

(7) Да се приложат адекватни статистически методи за анализ на взаимовръзките между разноименните индекси и параметри, определени в рамките на съответната теория.

(8) Да се приложат адекватни статистически методи за анализ на съгласуваността между едноименните индекси и параметри.

4. Методология на изследването

4.1. Дизайн

4.1.1. Променливи величини и статистически методи

Планираното изследване принадлежи към категорията на корелационните изследвания. Този тип изследвания се прилагат често при анализите в сферата на психологическите и образователни измервания - в случаите, в които условията за провеждане на експериментално изследване са неизпълними или неподходящи (Gribbons & Herman, 1997). Най-общо корелационните изследвания се фокусират върху оценката на силата и типа на взаимовръзката между две или повече променливи. Степента на взаимовръзка между величините, които представляват интерес, се определя чрез някой от коефициентите на корелация, подходящ за съответния тип данни.

Разкриването типа и структурата на тази взаимовръзка се счита за задължителна основа на по-нататъшния задълбочен анализ на данните (Калинов, 2010).

За да се изследва стабилността/ нестабилността на една статистика на тестовите въпроси, нейната независимост/ зависимост от конкретната извадка от и. л., е необходимо тази статистика да бъде наблюдавана поне двукратно, при едни и същи тестови въпроси, които са попаднали в различни тестови варианти, използвани в различни изпитни сесии. В този смисъл по-нататък в текста, когато говорим за тестови варианти, ще имаме предвид не толкова техните поредни номера, колкото обстоятелството, че те са използвани в различни тестови сесии, на които са се явили различни групи от кандидат-студенти.

Двукратното наблюдение на дадена статистика, направено върху всеки един от множество въпроси, използвани в два различни тестови варианта, предполага формирането на два вектора със стойности на тази статистика. Като мярка за степента на устойчивост, стабилност на статистиките на въпросите, на тяхната независимост от условията, при които са оценени, ще бъде използван коефициент на корелация, подходящ за типа на скалата, която формира съответната статистика. Параметрите на въпросите по *IRT* образуват интервална скала и поради това към тях ще бъде приложен Пиърсъновия коефициент на корелация между стойностите на съответната статистика, изчислени въз основа на резултатите от първата и втората извадка от и. л. За изследване на стабилността на индексите на въпросите в рамките на *CTT*, за които се предполага, че образува рангова скала (Fan, 1998; Анастази и Урбина, 2001), ще бъде приложен непараметричният коефициент на рангова корелация R на Спиърмън. Този коефициент се разглежда като специален случай на линейния коефициент на корелация на Пиърсън и предполага, че съответните променливи са измерени поне в порядкова скала (Калинов, 2010). Ще бъде направен опит за изследване на стабилността на тези индекси и чрез параметричен коефициент на корелация, като особено внимание ще бъде отделено на трудността на въпросите (p).

Разбира се, при това изследване статистическата функционалност на корелационното изследване ще бъде приложена в по-редуциран вид. Тук не може да се говори, в строгия смисъл на думата, за функционална

взаимовръзка между двете променливи, т.е. между двете оценки на съответната статистика. Психометричният смисъл на корелационния анализ е той да бъде в услуга на изследването на стабилността/ неизменяемостта/ съгласуваността на една или друга оценка на качеството на тестовите въпроси. Неговото приложение тук ще бъде ограничено до установяване на това дали има и каква е степента на съгласуваност между стойностите, получени при двете оценки. Поради качеството му, в което се използва в това изследване, корелационният коефициент следва да бъде разглеждан като „коефициент на стабилност" на статистиките на въпросите.

Имайки предвид границите на изменение на коефициентите на корелация, техните високи стойности, близки до 1.00, следва да бъдат интерпретирани като свидетелство за устойчивост на съответната статистика, за нейната независимост от условията на извадката. По-надолу ще бъдат фиксирани конкретни прагови стойности.

За верифициране на останалите допускания също ще бъдат приложени съответстващи на типа на променливите корелационни методи.

Поради характера на настоящото изследване в качеството си на променливи величини в него са включени следните статистики на въпросите:

(1) индекс на трудност (p)

(2) индекс на дискриминативна сила (D)

(3) бисериален коефициент на корелация (r_{bis}), определени в рамките на *CTT*

(4) параметър на дискриминативна сила (a)

(5) параметър на трудност (b)

(6) параметър на налучкване на правилния отговор (c), определени в рамките на *IRT*

4.1.2. Критерии за оценка на стабилността

Известно е, че коефициентът на Пиърсън е мярка за линейна корелация и може да достигне високи равнища не само тогава, когато стойностите на двете променливи по всеки обект (тестов въпрос) са равни или близки, но и тогава, когато стойностите на едната променлива са системно по-високи/ по-ниски от тези на другата. В този случай разпределенията на статистиките могат

да имат различни средни и, възможно, различни стандартни отклонения. Поради това корелационният коефициент на стабилност може да даде информация доколко наблюденията при второто измерване са възпроизвели относителните позиции на наблюденията в първото.

При променливите, измерени в рангова скала, може да се наблюдава същият ефект, ако стойностите на отделните скали принадлежат към различни подмножества от числовата система с отношения.

Статистиките на въпросите се използват като мярка за измерителните качества на съответния въпрос. Те са обект на внимателен анализ при процедурата на позитивен/ негативен подбор на въпросите след пилотното тестиране, за формиране на окончателния вариант на теста. Поради това увереността (ако има основания за такава увереност), че статистиките запазват относителните си позиции, е може би недостатъчна за признаване на тяхната стабилност. Например, ако бъде установена висока корелация между индексите за трудност на въпросите (p), това би могло да означава, че ако при един пилотен тест (първа извадка) даден въпрос j_1 има стойност p_{j1} = 0.43 и бъде оценен като качествен, то при един актуален тестов вариант (втора извадка) същият въпрос j_1 може да придобие стойност p_{j1} = 0.03 и да влоши общото качество на теста.

Ето защо като втори, съпътстващ метод за оценка на стабилността на тестовите статистики ще приемем съотношението между централните им тенденции при първото и второто измерване. Подходящ метод за оценка на такъв тип съотношения при променливи, измерени в интервална скала, е дисперсионния анализ с повторни измервания (*Repeated measures ANOVA*) с една независима променлива – вариант на теста (т.е. условие, при което е получена съответната статистика, което съответства на извадката от и. л., от резултатите на които е изчислена тази статистика) с две равнища, които условно ще обозначим като първа и втора оценка на съответната статистика. В този случай ще бъдат формулирани серия от нулеви хипотези за всяко сравнение с общ вид:

$$H_0 : \mu_1 = \mu_2$$

където: μ_1 и μ_2 – математически очаквания на стойностите на съответния

индекс/ параметър при първата и втората оценка

При неметричните скали като тази на индекса на дискриминативна сила по *CTT*, за който може да се предполага, че образува рангова скала, ще бъде приложен, подобно на корелационния анализ, по-подходящ статистически метод. Това е Знаково-ранговия тест на Уилкоксън за зависими извадки (*Wilcoxon matched pairs test*), който е непараметричен аналог на t-теста на Стюдънт за зависими извадки или на *ANOVA* с повторни измервания. Нулевата хипотеза, която подлежи на проверка е, че променливата, образувана от разликите ($d = x - y$) между всяка двойка стойности (x, y) има нулева медиана.

В по-схематична форма нулевата хипотеза може да се представи като равенство на медианите на разпределенията на двете изходни променливи:

$$H_0 : Me_x = Me_y$$

Съобразно двата подхода за изследване на стабилността на тестовите статистики, като критерий за оценка на тяхната вариативност/ инвариантност ще приемем следния конюнктивен модел, който включва две условия, които следва да бъдат удовлетворени едновременно:

(1) Стойности на коефициента на стабилност, равни на 0.70 или по-високи, ще бъдат интерпретирани като свидетелство за инвариантност на съответния индекс/ параметър.

Макар че корелационният анализ е може би най-широко използваният статистически метод в областта на психологическите изследвания, могат да бъдат приведени множество примери за различни интерпретации (оценки) на големината на получените коефициенти, което означава, че по този въпрос няма конвенция. Често цитирани са критериите (по-скоро – практически правила), предложени от Дж. Коен за оценка на големината на корелационния коефициент като мярка за размера (силата) на ефекта. Дж. Коен предлага коефициенти със стойност около 0.10 да се разглеждат като ниски, тези около 0.30 – като средни/ умерени и тези със стойности около 0.50 – като високи (Cohen, 1988, стр. 77–81). Авторът базира тези граници на обичайните, най-често наблюдавани стойности в при изследвания в областта на поведенческите науки и образователните измервания. Дж. Хемфил прави

интересен опит за разпростре тази класификация върху резултатите от корелационни изследвания в тези области, в които корелационните коефициенти се използват по обичайното им предназначение (Hemphill, 2003). Авторът анализира 380 мета-аналитични изследвания в областта на психологическите измервания и терапии, извличайки докладваните в тях корелационни коефициенти и размери на ефекта (последните също са трансформирани в корелационни коефициенти), сортира ги по възходящ ред на абсолютните стойности, след което ги разпределя в 3 последователни групи с приблизително еднакъв обем. Коефициентите в първата група като цяло са по-ниски от 0.20, във втората варират от 0.20 до 0.30, а в третата са над 0.30. Авторът предлага точно тези стойности като гранични, още повече, е намира много сходства между полученото от него емпирично разпределение на корелационните коефициенти и тези, получени при други изследвания, цитирани от него. Интересно е да се отбележи, че максималната стойност на коефициентите на корелация в горната третина, наблюдавани в изследвания в областта на психологическите измервания, е 0.78.

Макар че изглеждат твърде либерални, референтните стойности на Дж. Коен са по-скоро строги. Например стойността на r = 0.50 за голям размер на ефекта съответства на 89-тия процентил от разпределението на коефициентите на корелация в областта на психологическите измервания. Това означава, че 89% от получените корелационни коефициенти имат стойности, равни или по-малки от 0.50 (ibid.)

Въпреки това смятаме, че праговата стойност на коефициента на стабилност следва да бъде много по-консервативна. Приемаме праговата стойност от 0.70 по подобие на някои други мерки, основани на корелацията, като коефициента на стабилност при оценката на надеждността чрез повторно използване на един и същи тест или коефициента на надеждност на Кронбах.

(2) Като свидетелство за инвариантност на съответния индекс/ параметър ще бъдат разглеждани и случаите на потвърдена нулевата хипотеза за липса на разлика между средните стойности (*ANOVA* с повторни измервания) или медианите (при теста на Уилкоксън) при поне среден/ умерен размер на ефекта.

За оценка на размера на ефекта в случаите на повторни измервания се използва коефициентът на частна корелация ета на квадрат (*partial eta-*

squared), който отразява дела на вариацията на ефекта и на грешката в зависимите променливи, която може да бъде обяснена с въздействието на съответния фактор. За разлика от коефициента ета на квадрат, при коефициента на частна корелация няма общоприети норми за оценяване на размера на ефекта. Дж. Коен предлага практически правила, приложими главно към η^2, но които могат да се използват и при коефициента на частна корелация в случаите на еднофакторен дизайн: малък/ слаб – 0.01; среден/ умерен – 0.059; голям – 0.138 (Cohen, 1988).

4.2. Процедура за подбор на въпросите

Определянето на въпросите, отговарящи на посочените по-горе условия и подходящи за включване в емпиричното изследване, бе извършено с любезното съдействие на Центъра по оценяване към НБУ, който осигури достъп до резултатите от изпитите (до суровите данни) от наличните тестови варианти на ТОП от 1998 г. до 2008 г. Селекцията на въпросите бе извършена при спазване на изискванията на Центъра за конфиденциалност на информацията.

Първата задача при провеждане на емпиричното изследване бе да се идентифицират отделни въпроси (или групи от въпроси), които попадат в няколко (поне два) варианта на ТОП, използвани в различни тестови сесии.

Процедурата за подбор на въпроси, които отговарят на това условие, стартира с проучване на „паспортите" на отделните въпроси. Това са архивни записи, в които се съхранява и поддържа детайлна информация за тяхното администриране, включително и за тестовите варианти и изпитните сесии, в които са били използвани. След като бе установено наличието на такива въпроси, усилията бяха фокусирани върху това да се определи поне една, а при възможност – няколко по-големи групи от въпроси, всяка от които да е използвана при конструирането на два различни варианта, предназначени за различни изпитни сесии. След продължително проучване, в хода на което бяха отхвърлени множество въпроси, използвани в два или повече тестови варианта, но със слабо сечение помежду им, се откроиха три двойки от тестови варианти, в които се наблюдава почти пълно съответствие между въпросите

във всички раздели на ТОП.

Справка за избраните (двойки) тестови варианти е представена в следващата таблица. При първата и втората двойка броят на общите въпроси е 99 (99.00% от всички), а при третата – 100 (100%).

Таблица 1. Данни за тестовите варианти на ТОП, използвани в емпиричното изследване

Първа оценка				Втора оценка		
Номер на двойка варианти	**Вариант на ТОП**	**Дата на изпитна сесия**	**Брой изпитани**	**Кореспондира с вариант на ТОП**	**Дата на изпитна сесия**	**Брой изпитани**
1	92	20.04.2003	636	128	27.02.2005	543
2	96	18.04.2004	652	132	17.04.2005	638
3	110	27.06.2003	865	146	24.07.2005	454

Данните в таблицата свидетелстват, че тестовите варианти са използвани в различни изпитни сесии, при отделните двойки – в различни календарни години, с различен брой кандидат-студенти. Всичко това дава възможност, съгласно дизайна на изследването, за двукратна оценка на всяка статистика върху отделна, независима извадка.

4.3. Процедура за психометричен анализ на въпросите

След селектирането на двойките тестови варианти, суровите резултати от всеки вариант бяха обработени, по съответния ключ за правилните отговори, с програмния продукт Iteman, който е част от Item and Test Analysis Package, разработен от компанията Assessment Systems Corp. (Assessment Systems Corporation, 1997). Iteman е специализирана професионална програма за анализ на тестовите данни по Класическата тестова теория. Анализът бе извършен на ниво тест, т. е. при разглеждането на съответния тестова вариант като единна скала, включваща 100 въпроса. Сред различните резултати от приложението на алгоритмите на Класическата теория важни за настоящия анализ са числовите стойности на индексите на трудност (p) и дискриминативна сила (D) на въпросите от всеки тестов вариант.

След това суровите резултати от всеки тестов вариант бяха подложени и на процедура за калибриране, т. е. за оценка на параметрите на тестовите

въпроси съгласно трипараметричния модел на Теорията за отговор на тестов въпрос. Обработката на данните беше направена със софтуерния продукт Xcalibre, който е част от модула за анализ на айтеми и тестове от психометричния софтуер MicroCAT™ Testing System (ibid.) Програмата е специализирана за извършване на анализ по дву- и трипараметричния логистичен модел на *IRT*. За целите на тази част от изследването бе приложен трипараметричният модел, в рамките на който бяха изчислени стойностите на параметрите дискриминативна сила (a), трудност (b) и налучкване на правилния отговор (c).

За калибриране на параметрите се прилага метода на маргиналното правдоподобие (*Marginal maximum-likelihood, MML*). Този метод се състои в определянето на такива оценки на неизвестните параметри на въпросите и на индивида, които да максимизират съответните функции на правдоподобието. Чрез този метод се получават асимптотично ефективни оценки на параметрите.

В алгоритъма на Xcalibre методът *MML* за калибриране на параметрите на тестовите въпроси се реализира на няколко стъпки.

(1) Начална фаза, в която се прави предварителна оценка на параметрите, основана на техните индекси по *CTT* (трудност p, бисериален коефициент на корелация r_{bis} в качеството му на дискриминативен индекс и на налучкване на правилния отговор като реципрочна стойност на броя на алтернативните отговори). Стойностите на класическите индекси се трансформират по определен алгоритъм в първоначални оценки на параметрите a, b и c.

(2) Прилагане на алгоритъма *EM* (*Expectation - Maximization*), при който като начин за оценка на параметрите се използва максимизирането на функцията на максималното правдоподобие. Сам по себе си този алгоритъм е циклична двустъпкова процедура, която при стъпка „*E*" цели да се определи очаквания брой на лицата в популацията, разпределени между 15 предварително фиксирани точки на континуума Θ (в интервала от -3.5 до 3.5, през 0.50), и дела на онези лица във всяка група, които биха отговорили правилно на съответния въпрос. На стъпка „*M*", която е итеративна, се прави оценка на параметрите на всеки въпрос до удовлетворяване на предварително

определен критерий. Циклите *EM* за всички въпроси продължават до тогава, докато не бъде постигнат фиксирания критерий за толерантност и при най-„упоритията" въпрос. Този критерий представлява сумата от абсолютните стойности на измененията във всички параметри на даден въпрос при даден цикъл, в сравнение с предходния, и тази сума не трябва да надхвърля 0.05.

При оценката на параметрите се прилага Байесовия подход, съгласно който се предполага, че не само оценките, но и самите параметри са случайни величини, които имат някакво вероятностно разпределение. Плътността на разпределението на параметъра трябва да бъде известна преди да се направи неговата оценка, т. е. необходимо е да се определи някакво априорно разпределение на вероятностите. За да се подпомогне процеса на оценяване, за всеки параметър на въпросите в алгоритъма на програмата са определени различни априорни разпределения, с фиксирана средна стойност и стандартно отклонение.

4.4. Трансформиране на стойностите на индекса на трудност (*p*) в интервална скала

Една от обичайните процедури е нормализиране на данните, което се изразява в трансформирането на отделните стойности на изходното разпределение в z-единици на нормираното (стандартно) нормално разпределение. Л. Айкен отбелязва, че за разлика от ординалното измерване, „стандартизираните оценки представят измерването в интервална скала" (Aiken, 1988, стр. 87). Същата техника се препоръчва и използва от редица други изследователи (Fan, 1998; Анастази и Урбина, 2001).

Макар че, както бе отбелязано по-горе, една суб-интервална скала би могла да се третира като интервална, ние направихме допускането, че е възможно в скалата на трудността p да се наблюдава чувствително неравенство на интервалите и поради това е необходимо нейното нормализиране. За трансформацията на суровите стойности на индекса на трудността p в интервална скала бе приложена следната двустъпкова процедура.

(1) Преобразуване на стойностите на индекса p в процентили по формулата $P_i = (1 - p_i)$.

(2) Преобразуване на получените стойности в z-единици на стандартното нормално разпределение.

Тази процедура бе приложена към суровите стойности на индекса p, изчислени по съответния алгоритъм на СТТ за всички тестови варианти, обхванати в анализа. Тестовите варианти се разглеждат като единна скала, състояща се от 100 въпроса.

Първата стъпка се извършва поради това, че стойността на p съответства на относителния дял на лицата, които доминират над съответния въпрос съгласно теоретичния модел „Данни единичен стимул" на К. Кумбс (Coombs, 1964). Чрез това преобразуване се определя делът на лицата, които са доминирани от съответния въпрос, т.е. чиито идеални точки се намират наляво от неговата точка.

На втората стъпка, от статистическа таблица със стойностите на стандартното нормално разпределение, по получените процентилни стойности, които съответстват на частта от лицето на повърхнината под кривата на стандартното нормално разпределение от -∞ до дадената точка, се определя съответната z-стойност (Анастази и Урбина, 2001; Калинов, 2010). Преобразувани по този начин, стойностите на индекса p формират нормална крива и се изразяват в единиците на нормираното нормално разпределение със средна стойност M = 0.00 и стандартно отклонение SD = 1.00.

При анализа на стабилността на този индекс са използвани стандартизираните z-стойности, което позволява прилагането както на Пиърсъновия коефициент на корелация, така и на дисперсионния анализ с повторни измервания.

II. Резултати

1. Описателни статистики на зависимите променливи

Описателните статистики представят обобщена информация за разпределенията на зависимите променливи (индекси в рамките на Класическата тестова теория и параметри – на Теорията за отговор на тестов

въпрос). Те биха могли да дадат полезна информация за общото равнище на стойностите на съответната статистика, на тяхната хомогенност в рамките на анализираните тестови варианти, както и да послужат за първоначална оценка на качеството на тестовете като цяло. Всяка от статистики на въпросите са изчислени чрез алгоритмите на съответната теоретична рамка на описаните в теоретичната част тестови теории.

1.1. Характеристики на въпросите съгласно Класическата тестова теория

Ще започнем представянето на тестовите статистики с индексите на дискриминативна сила на въпросите, определени чрез алгоритмите на Класическата теория. Следващата таблица съдържа основните описателни статистики на неговото разпределение в шестте анализирани тестови варианта.

Таблица 2. Описателни статистики на дискриминативния индекс (*D*)

Двойка тестове	Вариант	Брой въпроси	Средна ст-т	Мин.	Макс.	Дисперсия	Станд. откл.
1.	вар. 92	100	0.201	-0.040	0.550	0.017	0.129
	вар. 128	99	0.203	-0.110	0.520	0.019	0.137
2.	вар. 96	100	0.208	-0.050	0.530	0.015	0.124
	вар. 132	99	0.216	-0.030	0.560	0.018	0.133
3.	вар. 146	100	0.193	-0.060	0.450	0.013	0.115
	вар. 110	100	0.206	-0.090	0.490	0.015	0.123

Първото нещо, което прави силно впечатление, са относително ниските средни стойности на този индекс, които се наблюдават във всички тестови варианти. Максималната средна в горната таблица е тази при вариант 132 (0.216), а минималната - при вариант 92 (0.201). Сравнително ниските средни стойности на индекса на дискриминативна сила, които при различните тестови варианти са устойчиво стабилизирани в диапазона от 0.20 до 0.22, са свидетелство, че тестовите варианти като инструменти на измерване не разграничават добре лицата с по-ниски от тези с по-високи способности. Това обстоятелство е тревожно и от прагматична гледна точка, тъй като, съгласно едно установено практическо правило, минималната стойност на индекса D, която се приема за долна граница на приемливост, е точно 0.20 (Ebel, 1954).

Макар че всички варианти съдържат въпроси с негативни стойности на този индекс, добрата новина е, че са малко на брой (около 4% – 6% от въпросите във всеки вариант), а и отклоненията наляво от нулевата стойност са слаби. По-съществени са те при вариант 128, който включва въпроси с най-ниска стойност на този индекс (-0.110). От друга страна, най-високите стойности на този индекс са в диапазона 0.45 до 0.55, което е далеч от горната граница на неговото изменение.

Сходството между минималните и максималните стойности на индекса в отделните тестови варианти намира отражение в приблизително еднаквите оценки на тяхното разсейване. Стойностите на стандартните отклонения са в диапазона от 0.115 при вариант 146 до 0.137 при вариант 128. Независимо от това, че стойностите на този индекс едва прекрачват зоната отвъд нулевата стойност и не надхвърлят умерените равнища от 0.50 -0.60, степента на тяхното разсейване едва ли може да бъде пренебрегната. С други думи, дискриминативната сила на въпросите е вариативен признак, по който те се различават и който следва да бъде включен при тяхното моделиране в *IRT*.

Бисериалният коефициент на корелация е втора, статистическа мярка за разграничителната способност на въпросите. По същество двата индекса (заедно с „класическия" индекс) са източник на една и съща информация,още повече, че имат едни и същи граници на изменение. Практиката показва, че бисериалният коефициент е по-консервативен от класическия. Това се потвърждава и от данните в следващата таблица.

Таблица 3. Описателни статистики на бисериалния коефициент на корелация (r_{bis})

Двойка тестове	Вариант	Брой въпроси	Средна ст-т	Мин.	Макс.	Дисперсия	Станд. откл.
1.	вар. 92	100	0.199	-0.270	0.500	0.021	0.144
	вар. 128	99	0.207	-0.240	0.560	0.026	0.162
2.	вар. 96	100	0.233	-0.210	0.600	0.023	0.150
	вар. 132	99	0.233	-0.070	0.520	0.021	0.145
3.	вар. 146	100	0.188	-0.180	0.610	0.017	0.131
	вар. 110	100	0.198	-0.220	0.520	0.017	0.130

Средната стойност на този индекс се мени в сравнително тесни граници, като максималната стойност от 0.233 се наблюдава при два тестови варианта

(96 и 132), минималната му стойност е 0.188 при вариант 146. Като цяло средните стойности на този статистически индекс на различителната сила са малко по-ниски от тези на класическия.

Забелязват се обаче значително по-ниските минимални стойности, които при всички тестови варианти имат отрицателен знак, за да достигнат до -0.270 при вариант 92. От друга страна, максималните стойности са малко по-високи в сравнение с тези от предходната таблица, достигащи до 0.610 при тестов вариант 146. Тази по-широк диапазон на изменение на бисериалния коефициент намира израз в по-високите стойности на стандартните отклонения в сравнение с тези при класическия индекс. Както би могло да се очаква, анализът на бисериалния коефициент като втора мярка на разграничителната способност на въпросите потвърждава направените по-горе оценки, че като цяло тестовите въпроси не различават добре лицата от контрастните групи и още, че тази характеристика на въпросите е вариативен признак, по който те се различават и който следва да бъде включен при тяхното моделиране в *IRT*.

Трудността на въпросите е характеристика, която влияе в най-висока степен на тестовите резултати, а и на трудността на теста като цяло. Данните от следващата таблица показват, че отделните тестови варианти се характеризират с приблизително еднаква средната трудност на въпросите, която се изменя в диапазона от 0.41 до 0.45. Поставена в контекста на теоретичните граничните стойности на този индекс, трудността на въпросите от различните варианти следва да се разглежда като превишаваща средното равнище.

Таблица 4. Описателни статистики на индекса на трудност (*p*)

Двойка тестове	Вариант	Брой въпроси	Средна ст-т	Мин.	Макс.	Дисперсия	Станд. откл.
1.	вар. 92	100	0.432	0.030	0.980	0.051	0.225
	вар. 128	99	0.417	0.030	0.980	0.050	0.224
2.	вар. 96	100	0.439	0.040	0.970	0.060	0.248
	вар. 132	99	0.449	0.070	0.970	0.058	0.242
3.	вар. 146	100	0.408	0.060	0.980	0.051	0.226
	вар. 110	100	0.409	0.060	0.990	0.052	0.227

За разлика от индексите на различителна сила, трудността варира в много широк диапазон. Минималните стойности при всички варианти се

приближават към долната граница на изменение на този индекс, достигайки до 0.030 (при варианти 92 и 128), максималните – към горната граница на изменение, достигайки до 0.990 при вариант 110.

Сходните средни, минимални и максимални стойности на този индекс намират отражение в относително високите (заради големия размах), но приблизително еднакви стойности на стандартните отклонения. Следователно, съдейки по тези основни характеристики на разпределенията на индексите на трудност, можем да заключим, че в отделните тестови сесии, проведени през различни години, трудността на вариантите се проявява като тяхна устойчива характеристика, която не се изменя съществено.

1.2. Характеристики на въпросите съгласно Теорията за отговор на тестов въпрос

Числовите стойности (оценките) на параметрите, изчислени въз основа на алгоритмите на трипараметричния модел на Теорията за отговор на тестов въпрос, са получени в рамките на съвършено различен психометричен модел. Поради това описателните статистики на техните разпределения не могат да бъдат пряко съпоставени, но ние ще очертаем някои паралели, основани на възможните (или типичните) граници на техните изменения, както и на наблюдаваното равнище на хомогенност р рамките на съответните тестови варианти.

В следващите няколко таблици са представени основните описателни статистики на разпределенията на параметрите на въпросите в различните тестови варианти, предмет на настоящото изследване.

Средните стойности на дискриминативния параметър (a) при различните варианти се движат в сравнително тесните граници между 0.507 (при варианти 92 и 146) и 0.594 (при вариант 132). Интересно е, че се наблюдава известна разлика в средните равнища на този параметър между двойките варианти, съставени от едни и същи въпроси. Без да надценяваме тези наблюдения (поради липса на данни за значимостта на тези разлики), ще отбележим, че те могат да бъдат сигнал за определена вариативност на стойностите на този параметър.

Таблица 5. Описателни статистики на дискриминативния параметър (*a*)

Двойка тестове	Вариант	Брой въпроси	Средна ст-т	Мин.	Макс.	Дисперсия	Станд. откл.
1.	вар. 92	100	0.507	0.320	0.860	0.012	0.110
	вар. 128	99	0.554	0.340	0.850	0.013	0.114
2.	вар. 96	100	0.574	0.360	0.890	0.013	0.112
	вар. 132	99	0.594	0.370	0.860	0.013	0.114
3.	вар. 146	100	0.507	0.350	0.740	0.008	0.089
	вар. 110	100	0.525	0.300	0.900	0.015	0.124

Друга особеност, която следва да бъде посочена, е липсата на въпроси с негативни стойности по този параметър, за разлика от едноименния индекс, определен по *CTT*. Минималните наблюдавани стойности на въпросите при всички тестови варианти са положителни, като най-ниската сред тях е 0.300 (при вариант 110), а най-високата достига 0.900 (при същия вариант).

Съпоставени с едноименните индекси по Класическата теория, дискриминативните параметри са малко по-хомогенни, с по-слаба вариация. При по-голяма част от анализираните тестови варианти стандартното отклонение варира около 0.11 - 0.12, при 0.12 – 0.14 за класическия индекс D и 0.13 - 0.16 за бисериалния коефициент на корелация r_{bis}. С други думи, дискриминативният параметър на въпросите варира в степен, която е съпоставима с тази на индексите от класическата теория и поради това включването му в моделирането на въпросите изглежда оправдано.

Подобно на данните за класическия индекс на трудност, средните стойности на едноименния параметър свидетелстват за това, че въпросите се отличават с повишена трудност (в границите на изменение от ±3.00, наложени от използвания софтуер), която варира от 1.102 (при вариант 132) до 1.401 (при вариант 110).

Таблица 6. Описателни статистики на параметъра трудност (*b*)

Двойка тестове	Вариант	Брой въпроси	Средна ст-т	Мин.	Макс.	Дисперсия	Станд. откл.
1.	вар. 92	100	1.280	-3.000	3.000	2.948	1.717
	вар. 128	99	1.271	-3.000	3.000	2.638	1.624
2.	вар. 96	100	1.122	-3.000	3.000	3.398	1.843
	вар. 132	99	1.102	-3.000	3.000	3.250	1.803
3.	вар. 146	100	1.386	-3.000	3.000	2.871	1.694
	вар. 110	100	1.401	-3.000	3.000	2.760	1.661

Тук също се забелязват различия между средната трудност в рамките на отделните двойки варианти, което би могло да бъде индикация за вариативност на този параметър.

Друго сходство с класическия индекс може да бъде открито в широките граници на изменение на стойностите на параметъра, които покриват целия посочен по-горе интервал. Това намира отражение във високите стойности на стандартното отклонение, които обаче надвишават многократно тези при класическия индекс. Обяснение за тази разлика може да бъде намерено в разпределенията на стойностите на съответните статистики. Докато разпределенията на класическия индекс се приближават към нормалното (с положителна асиметрия), то разпределенията на съответния параметър са по-скоро L-образни, с отрицателна асиметрия и с ярко изразен таванен ефект. Източникът на високите стандартни отклонение е натрупването на високи честоти в десния край на разпределенията.

Статистиките на разпределенията на параметъра на склонността към налучкване на правилните отговори са интересни преди всичко с това, че съдържат свидетелства за жизнеността на тази характеристика на въпросите. Средните стойности на параметъра в различните тестови варианти са близки, около 0.17 – 0.19, което е малко под очакваната стойност от 0.20 (1/5, съобразно броя на дистракторите във всички въпроси.

Таблица 7. Описателни статистики на параметъра склонност към налучкване (*c*)

Двойка тестове	Вариант	Брой въпроси	Средна ст-т	Мин.	Макс.	Дисперсия	Станд. откл.
1.	вар. 92	100	0.176	0.090	0.210	0.000	0.021
	вар. 128	99	0.178	0.100	0.210	0.000	0.021
2.	вар. 96	100	0.188	0.100	0.260	0.001	0.026
	вар. 132	99	0.192	0.110	0.240	0.001	0.023
3.	вар. 146	100	0.169	0.110	0.190	0.000	0.016
	вар. 110	100	0.173	0.100	0.210	0.000	0.022

Интересно е, че максималните стойности надвишават слабо средните, което очертава друга особеност на разпределенията на този параметър. Те са силно асиметрични, скосени отдясно, с концентрация на високи честоти непосредствено около средната стойност и ниски, намаляващи честоти под

стойности от 1.15 – 0.16. Тази особеност намира отражение и в статистиките на разсейването, които имат много ниски, приблизително еднакви стойности. Всичко това показва, че склонността към налучкване е устойчива характеристика на въпросите, която, заедно с тяхната дискриминативна сила, не трябва да бъде пренебрегвана при тяхното описание.

2. Анализ на стабилността на статистиките на въпросите

2.1. Стабилност на индексите на въпросите, определени в съответствие с Класическата тестова теория

Както беше отбелязано, важно качество на тестовите въпроси е стабилността на техните индекси. Ако техните стойности се изменят в зависимост от тестовата сесия, в която са използвани, съответно от извадката, въз основа на която са изчислени, това не само би подронило доверието в тези индекси като характеристики на въпросите, но би възпрепятствало и създаването на еквивалентни тестове.

Поради ранговия характер на тези индекси, като първа оценка на тяхната стабилност са използвани коефициентите на рангова корелация R на Спиърмън и, в допълнение, на линейна корелация r_{xy} на Пиърсън, приложени върху две серии от стойности на съответния индекс, получени от едни и същи съвкупности от тестови въпроси, използвани в две различни условия, описани по-горе. Ако при това съпоставяне се наблюдават високи равнища на корелация и по-конкретно – по-високи от приетата прагова стойност, то това би означавало, че съответната статистика е стабилна и не се влияе от извадката, въз основа на която е изчислена.

В допълнение, като втора мярка на стабилността на индексите е направена статистическа оценка на разликата между техните средни стойности или медиани. Ако нулевата хипотеза за липса на такава разлика не може да бъде отхвърлена и при наличие на висока корелация, това би означавало, индексите на съответните въпроси съхраняват не само относителните, но и абсолютните си позиции на съответна скала.

Дискриминативна сила (D)

В следващата таблица са представени резултатите от корелационния анализ на индексите на дискриминативна сила на въпросите, изчислени по CTT за съответните двойки тестови варианти. Както беше отбелязано, поради ранговия характер на индекса, като мярка за неговата стабилност е използван коефициентът на рангова корелация R на Спиърмън. Успоредно с това са изчислени и съответните коефициенти на линейна корелация. В последната колона е представена статистическата значимост на получените корелационни коефициенти.

Таблица 8. Коефициенти на стабилност (*R* и r_{xy}) на индекса на дискриминативна сила *D* в рамките на CTT

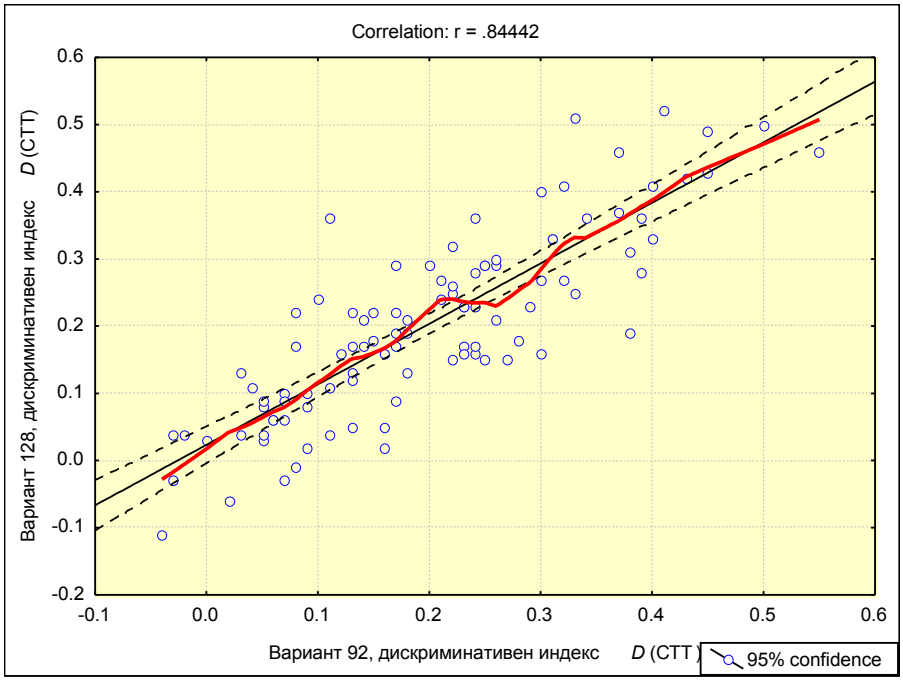
Двойка тестове	Тестов вариант	Коефициенти на стабилност (*R*, r_{xy})	Статистическа значимост (*p*)
1	вар. 92 вар. 128	R = 0.827 r_{xy} = 0.844	p < 0.05 p < 0.05
2	вар. 96 вар. 132	R = 0.820 r_{xy} = 0.819	p < 0.05 p < 0.05
3	вар. 146 вар. 110	R = 0.796 r_{xy} = 0.817	p < 0.05 p < 0.05

Данните в таблицата се характеризират с някои интересни особености. Коефициентите на стабилност на дискриминативната сила (R на Спиърмън) при трите двойки тестови варианта са положителни, високи и статистически значими на ниво $p < 0.05$. Те се намират в сравнително тесния интервал от 0.80 до 0.83, който е разположен далеч над приетата прагова стойност от 0.70. Същевременно наблюдаваните стойности са близки по големина, което дава основание да се каже, че устойчивостта не е единично явление. Интересно е да се отбележи, че равнищата на двата типа корелационни коефициенти – на рангова и на линейна корелация, при всяка двойка варианти са твърде близки. При една от двойките тестови варианти (96 – 132) стойността на ранговия коефициент е по-висока от тази на линейния, но при останалите два от случаите (вар. 92 – вар. 128 и вар. 146 – вар. 110) коефициентите на рангова корелация дори са малко по-ниски тези на линейна корелация. Следователно във взаимовръзката между тези индекси се наблюдава ясно изразен линеен

компонент.

Свидетелство за линейния характер на взаимовръзката между индексите на дискриминативна сила е следващата диаграма на разсейването на техните стойности от двойката варианти 92 и 128.

Фигура 1. Разсейване на индексите на дискриминативна сила *D* на въпросите от варианти 92 и 128



За апроксимиране на точките са приложени два модела – линеен и нелинеен. Приложеният нелинеен регресионния модел е известен като локално претеглена регресия (*locally weighted scatterplot smoothing*, LOWESS). Тя се определя за всяка точка (обект) и за най-близките до него точки. Приема се, че този метод води до по-добро представяне на формата на връзката между съответните две променливи. Функцията, свързваща индексите на дискриминативна сила на въпросите от двата тестови варианта, определена по този метод, може да се определи като монотонно нарастваща. В един от сегментите (в интервала 0.22 -0.26) по хоризонталната ос се наблюдава дори обратна тенденция, към понижаване на стойностите по вертикалната ос с увеличаване на тези по хоризонталната.

Като цяло обаче на горната графика може да се забележи ясно изразения линеен характер на взаимовръзката между стойностите на D на

въпросите в двата теста. Той се изразява във формата и ориентацията на облака от точки, представящи въпросите, както и от сравнително слабото им разсейване. Въпросите са групирани около регресионната линия, макар че има само някои отделни въпроси, чиито координати ги поставят извън общата маса на останалите. Така например над регресионната линия се открояват два въпроса с номера 24 и 88 с координати (стойности на дискриминативния индекс) съответно (0,11; 0,36) и (0,33; 0.51. Под линията като изключение се очертава въпрос 15 със стойности (0,38; 0,19). Отстраняването само на тези три въпроса би повишило стойността на R от 0.827 на 0.851, а на r_{xy} – от 0.844 на 0.875. Следва да обърнем внимание и на още една особеност. В два от сегментите на графиката на монотонната функция, в интервалите -0.04 – 0.17 и 0.31 – 0.55, нейната форма е изгладена и почти съвпада с линейната регресионна права.

Ще направим още една оценка на формата на връзката между двете оценки на D чрез проверка на хипотезата за нулева стойност на регресионния коефициент b:

$$H_0 : b = 0.00$$

Таблица 9. Оценка на регресионния коефициент

	b	**Стандартна грешка на b**	***B***	**стандартна грешка на B**	***t* (97)**	***p***
свободен член			0.023	0.0138	1.699	0.092
D (вар. 92)	*0.844	*0.054	*0.901	*0.058	*15.525	*0.000

Нулевата хипотеза може да бъде отхвърлена на ниво α = 0.05. Данните сочат не само за високата стойност на ъгловия коефициент, но и за ниската стандартна грешка на неговата оценка. Цялостната оценката на годността на модела, направена чрез *ANOVA*, е също висока – $F(1, 97) = 241.028$, $p = 0.000$. Изравненият коефициент на детерминация R^2 квадрат има стойност 0.710. Той отразява онази част от дисперсията на зависимата променлива в популацията от въпроси, която може да бъде обяснена чрез построения линеен модел. наблюдаваната стойност също е свидетелство за качеството на този модел.

Като втора мярка на стабилността на индекса на дискриминативна сила D е направена статистическа оценка на разликата между медианите на стойностите на въпросите от съответните два тестови варианта. Нулевата хипотеза е проверена чрез Знаково-ранговия T-тест на Уилкоксън за зависими

извадки (*Wilcoxon matched pairs test*). На следващата таблица са представени резултатите от теста при отделните двойки варианти на ТОП.

Таблица 10. Резултати от Знаково-ранговия тест на Уилкоксън за индекса на дискриминативна сила *D*, изчислен по СТТ

Двойка тестове	Тестов вариант	Брой наблюдения	Медиана	Тестова статистика (*T*)	Статистическа значимост (*p*)
1	вар. 92 вар. 128	99	0.190 0.190	1862.000	0.690
2	вар. 96 вар. 132	99	0.205 0.210	1848.000	0.780
3	вар. 146 вар. 110	100	0.180 0.190	1790.000	0.095

Съдейки по стойностите на тестовата статистика T и нейната статистическа значимост p, няма основания за отхвърляне на нулевата хипотеза при нито една от двойките тестови варианти. Тестът на Уилкоксън показва, че поставянето на тестовите въпроси в различни условия не предизвиква статистически значими разлики в средните равнища на техните дискриминативни индекси. Може да се приеме, че медианите на разпределенията на този индекс в съответните двойки тестови варианти са равни. Действително, дори и като извадкови статистики те имат равни (при вар. 92 и вар. 128) или много близки стойности (при останалите двойки тестови варианти).

Бисериален коефициент на корелация (r_{bis})

В рамките на *CTT* бисериалният коефициент на корелация r_{bis} се използва като втора мярка на дискриминативната сила на въпросите, наред с класическия дискриминативен индекс. За оценка на неговата стабилност са използвани същите коефициенти на рангова и на линейна корелация. Резултатите от анализа са представени в следващата таблица.

В таблицата също се наблюдават високи коефициенти на стабилност, които надвишават установената прагова стойност от 0.70. Сравнени със стойностите от предходния анализ на класическия дискриминативен индекс D, те се характеризират с малко по-ниски равнища. От друга страна, стойностите

попадат в сравнително тесния интервал от 0.78 до 0.81, което говори за възпроизводимостта на бисериалния коефициент като мярка на дискриминативната сила на въпросите.

Таблица 11. Коефициенти на стабилност (*R* и r_{xy}) на бисериалния коефициент на корелация r_{bis} в рамките на CTT

Двойка тестове	Тестов вариант	Коефициенти на стабилност (*R*, r_{xy})	Статистическа значимост (*p*)
1	вар. 92 вар. 128	R = 0.792 r_{xy} = 0.811	p < 0.05 p < 0.05
2	вар. 96 вар. 132	R = 0.808 r_{xy} = 0.797	p < 0.05 p < 0.05
3	вар. 146 вар. 110	R = 0.728 r_{xy} = 0.779	p < 0.05 p < 0.05

Вторият критерий за стабилност на бисериалния коефициент като мярка на дискриминативната сила на въпросите се основава на статистическата оценка на разликата между медианите на съответните двойки тестови варианти, верифицирана чрез Знаково-ранговия T-тест на Уилкоксън за зависими извадки. Резултатите от анализа са представени в следващата таблица.

Таблица 12. Резултати от Знаково-ранговия тест на Уилкоксън с повторни измервания за бисериалния коефициент на корелация r_{bis} в рамките на *CTT*

Двойка тестове	Тестов вариант	Брой наблюдения	Медиана	Тестова статистика (*T*)	Статистическа значимост (*p*)
1	вар. 92 вар. 128	99	0.215 0.220	2046.500	0.304
2	вар. 96 вар. 132	99	0.230 0.230	2148.000	0.624
3	вар. 146 вар. 110	100	0.180 0.195	2113.500	0.537

Съдейки по тестовите статистики и асоциираните с тях равнища на статистическа значимост, нулевата хипотеза за равенство на медианите не може да бъде отхвърлена при нито една от анализираните двойки тестови варианти.

Трудност (p)

Трудността на въпросите p е характеристика, която пряко повлиява трудността на целия тест. Както беше отбелязано, в резултат на начина на неговото изчисляване в рамките на *CTT*, този индекс на въпросите формира рангова скала. Поради това наблюдаваните „сурови" стойности на p бяха трансформирани в z-единици на нормираното нормално разпределение, формиращи интервална скала. Предишни наши изследвания обаче дават основание да се предположи, че скалата на индекса на трудността p не е ординална в строгия Стивънсов смисъл на понятието че тя притежава „надрангови" характеристики (Stevens, 1939, 1946; Стивенс, 1960). В по-общ методологически план редица изследователи поддържат тезата, че ранговите скали могат да бъдат третирани като интервални, без това да повлияе значително върху значимостта на направените изводи (Abelson & Tukey, 1959; Gaito, 1960; Baker, Hardyck & Petrinovich, 1966; McNemar, 1969; Gardner, 1975).

Ето защо при изследването на стабилността на индекса на трудността са приложени три различни подхода. Първо, скалата на трудността p е разгледана като строго ординална и за оценка на стабилността на индексите е използван коефициентът на рангова корелация R на Спиърмън. Второ, тя е третирана като интервална и за оценка на стабилността е използван коефициентът на линейна корелация r_{xy} на Пиърсън. И накрая, като основа за оценката на стабилността са използвани трансформираните в z-единици сурови стойности, върху които също е приложен коефициентът на линейна корелация.

Таблица 13. Коефициенти на стабилност (*R* и r_{xy}) на индекса на трудност (*p)* в рамките на CTT

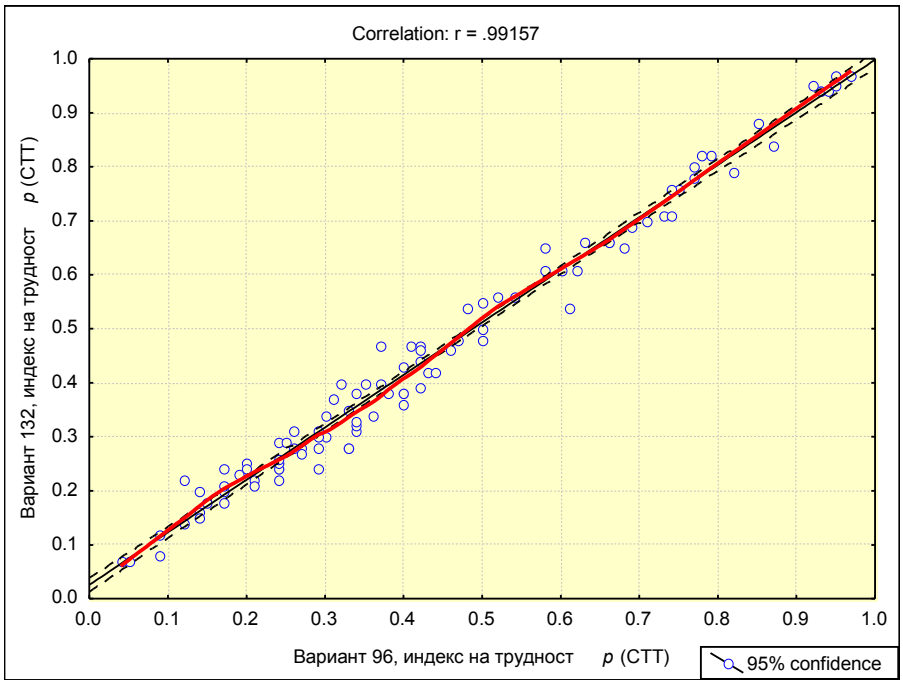
Двойка тестове	Тестов вариант	Коефициенти на стабилност (*R*, r_{xy})	Статистическа значимост (*p*)
1	вар. 92 вар. 128	R (p) = 0.924 r_{xy} (p) = 0.928 r_{xy} z(p) = 0.930	p < 0.05 p < 0.05 p < 0.05
2	вар. 96 вар. 132	R (p) = 0.986 r_{xy} (p) = 0.992 r_{xy} z(p) = 0.990	p < 0.05 p < 0.05 p < 0.05
3	вар. 146 вар. 110	R (p) = 0.978 r_{xy} (p) = 0.986 r_{xy} z(p) = 0.985	p < 0.05 p < 0.05 p < 0.05

Всички корелационни коефициенти имат стойности над 0.90, в повечето случаи достигат 0.98 или 0.99 при нива на статистическа значимост под 0.05. Може да се отбележи, че наблюдаваните стойности на коефициента на рангова корелация са малко по-ниски от тези на линейна корелация, но разликите между тях са несъществени, едва във втория десетичен знак.

Коефициентът на стабилност, чиято валидност е под въпрос – коефициентът на линейна корелация, приложен върху суровите стойности на p, образуващи рангова скала, достига същите равнища, както и предходните два. Нещо повече, най-високата стойност сред всички коефициенти на стабилност (0,992, при двойката варианти 96 - 132) се наблюдава именно при r_{xy}, приложен върху суровия индекс p.

Преди да разгледаме особеностите на връзката между суровите стойности на p, ще добавим, че се наблюдава известна разлика между равнищата на коефициентите на стабилност в отделните двойки тестови варианти. При първата двойка стойностите на всички статистики гравитират около 0.93, а при втората и третата – около 0.98 – 0.99. Това е още едно свидетелство за взаимната заменяемост на тези мерки на стабилността.

Фигура 2. Разсейване на суровите стойности на индекса на трудност (*p*) на въпросите от варианти 96 и 132



Диаграмата на горната графика илюстрира линейния характер на взаимовръзката между суровите (рангови) стойности на p от два тестови варианта. Моделът е апроксимиран спрямо данните чрез стандартната линейна функция от типа $y = a + bx$, както и чрез локално претеглена регресия, използвана и за анализ на взаимовръзките между индексите D.

На тази диаграма може лесно да се забележи, че точките въпросите са разположени на (или много близо до) регресионната линия (представена с непрекъсната линия в черно). Почти липсват въпроси, чийто точки да се отклонява значително от този модел. Като вземем предвид и наклона на регресионната линия, може да заключим, че е разположена под ъгъл, който предполага много близки, почти съвпадащи стойности на индекса на трудност при двата теста. Забелязва се известно натрупване на точки в интервала 0.15 – 0.45, което съответства на големия брой въпроси с индекси на трудност в този интервал.

Функцията, която описва съпоставянето на суровите стойности на трудността p от двата варианта на теста, определена по метода *LOWESS*, може да се разглежда като монотонно растяща. Нейното графично представяне е вълнообразна линия (представена с плътна непрекъсната линия в червено). Трябва да признаем обаче, че монотонната функция следва почти дословно регресионната права на линейната функция и че нейните отклонения в едната или другата посока са минимални. Забелязва се още, че тази функция е (почти) линейна в интервала 0.60 – 1.00.

Втората оценка на стабилността на индекса на трудност е направена върху стандартизираните z-стойности, с прилагане на *ANOVA* с повторни измервания.

Таблица 14. Резултати от дисперсионния анализ с повторни измервания за стандартизирания индекс на трудност *z*(p) в рамките на CTT

Двойка тестове	Тестов вариант	Сума от квадратите	Степени на свобода	Среден квадрат	Тестова статистика (*F*)	Статист. знач. (*p*)
1	вар. 92 вар. 128	0.094	1	0.094	2.769	0.099
2	вар. 96 вар. 132	0.132	1	0.132	23.587	0.000
3	вар. 146 вар. 110	0.000	1	0.000	0.018	0.894

Резултатите от последователните тестове на нулевата хипотеза за отделните двойки варианти не са консистентни. При първата и третата двойка нулевата хипотеза не може да бъде отхвърлена, но при втората тя може да бъде отхвърлена при равнище на p = 0.00. Следователно, средните стойности на стандартизираните индекси на вариантите 96 и 132 не са равни, макар че наблюдаваната разлика между тях (съответно 0.166 и 0.115) е само 0.052 единици.

Дали тази разлика е голяма? За да се даде отговор на този въпрос, следва да се направи оценка на размера на ефекта, което е и изискване на APA (Wilkinson, L., & APA Task Force on Statistical Inference, 1999). За оценка на размера на ефекта в случаите на повторни измервания се използва коефициентът на частна корелация ета на квадрат (*partial eta-squared*), който отразява дела на вариацията на ефекта и на грешката в зависимите променливи, която може да бъде обяснена с въздействието на съответния фактор. Взаимодействието между двата типа променливи може да се разглежда като „корелация" между тях, поради което чрез степенуването на този коефициент може да се определи „чистият" ефект.

При втората двойка размерът на ефекта е $\eta_p^2 = 0.194$, при мощност на критерия 0.998 (при α = 0.05). С други думи, 19.4% от дисперсията в стандартизираните стойности на индекса на трудност може да бъде обяснена с обстоятелството, че те са определени въз основа на различни извадки от и. л. Ако се придържаме към критериите на Дж. Коен за оценка на големината на ефекта, определен чрез коефициента на частна корелация η^2, следва да признаем, че ефектът на извадките от и. л. върху трудността на въпросите при тази двойка тестови варианти е твърде съществен.

2.2. Стабилност на параметрите на въпросите, определени в съответствие с Теорията за отговор на тестов въпрос

Дискриминативна сила (a)

Методите, използваните за анализ на стабилността на параметрите на

тестовите въпроси, определени в рамките на *IRT*, се определят от интервалния характер на скалите, образувани от тях. Това са коефициентът на линейна корелация на Пиърсън и дисперсионният анализ с повторни измервания.

В следващата таблица са представени резултатите от корелационния анализ за избраните три двойки тестови варианти.

Таблица 15. Коефициенти на стабилност на параметъра на дискриминативна сила (*a*) в рамките на *IRT*

Двойка тестове	Тестов вариант	Коефициент на стабилност (r_{xy})	Статистическа значимост (*p*)
1	вар. 92 вар. 128	0.901	p < 0.05
2	вар. 96 вар. 132	0.888	p < 0.05
3	вар. 146 вар. 110	0.854	p < 0.05

Съответствията между стойностите на този параметър в двойките тестови варианти са много високи, както става ясно от получените коефициенти на корелация. Те варират в границите 0.85 до 0.90 и са статистически значими на ниво $p < 0.05$. Макар че равнищата на стабилност при отделните двойки варианти се различават, всички те са над приетата прагова стойност от 0.70 и показват високата степен на съгласуваност на параметрите на дискриминативна сила.

Изненадващи обаче са резултатите от проверката на съотношенията между средните стойности на този параметър. Данните в следващата таблица дават основание нулевата хипотеза да бъде последователно отхвърлена на равнища на $p < 0.05$.

Таблица 16. Резултати от дисперсионния анализ с повторни измервания на параметъра на дискриминативна сила (*a*) в рамките на *IRT*

Двойка тестове	Тестов вариант	Сума от квадратите	Степени на свобода	Среден квадрат	Тестова статистика (*F*)	Статист. знач. (*p*)
1	вар. 92 вар. 128	0.111	1	0.111	88.725	0.000
2	вар. 96 вар. 132	0.017	1	0.017	12.068	0.001
3	вар. 146 вар. 110	0.017	1	0.017	7.443	0.007

Това означава, че като цяло стойностите на този параметър в едната скала са изместени спрямо тези от другата със стъпка, равна на разликата между средните им стойности.

Нека да разгледаме още една група от статистики, чрез които можем да направим допълнителна оценка на получените резултати.

При първата двойка (вар. 92 – 128) разликата между наблюдаваните средни стойности е 0.047, но това води до размер на ефекта $\eta_p^2 = 0.475$, при мощност на критерия 1.000 (при α = 0.05). С други думи, 47.5 % от дисперсията в стойностите на този параметър може да бъде обяснена с обстоятелството, че те са определени въз основа на различни извадки от и. л.

При втората двойка ((вар. 96 – 132) разликата между наблюдаваните средни стойности е 0.019, което води до по-малък размер на ефекта $\eta_p^2 = 0.110$, при мощност на критерия 0.931 (при α = 0.05). Поради това 11.0 % от дисперсията в стойностите на този параметър може да бъде обяснена с обстоятелството, че те са определени въз основа на различни извадки от и. л.

При третата двойка (вар. 110 – 146) разликата между наблюдаваните средни стойности е по-малка от предходната (0.018) и поради това размерът на ефекта е по-малък - $\eta_p^2 = 0.070$, при мощност на критерия 0.771 (при α = 0.05). С други думи, 7.0 % от дисперсията в стойностите на този параметър може да бъде обяснена с обстоятелството, че те са определени въз основа на различни извадки от и. л.

Съгласно критериите на Дж. Коен, при първата двойка размерът на ефекта е голям, при втората – по-скоро умерено голям, а при третата - среден.

Трудност (b)

Устойчивостта на параметъра трудност на въпросите е на малко по-високо равнище, отколкото тяхната дискриминативна сила. Данните от анализа на този параметър са представени в следващата таблица. Данните показват каква е взаимовръзката между трудността на въпросите в различните двойки тестови варианти.

Таблица 17. Коефициенти на стабилност (r_{xy}) на параметъра трудност (*b*) в рамките на *IRT*

Двойка тестове	Тестов вариант	Коефициент на стабилност (r_{xy})	Статистическа значимост (*p*)
1	вар. 92 вар. 128	0.967	p < 0.05
2	вар. 96 вар. 132	0.972	p < 0.05
3	вар. 146 вар. 110	0.915	p < 0.05

Тук се наблюдават най-високите коефициенти на стабилност от всички, изчислени чрез теоретичния модел на Теорията за отговор на тестов въпрос, със статистическа значимост на равнище $p < 0.05$. От представените данни се вижда, че равнищата на устойчивост на трудността при различните двойки тестови варианти, макар и различни, се намират д интервала 0.92 – 0.97 и следователно надхвърлят праговата стойност от 0.70.

В допълнение, резултатите от дисперсионния анализ, представени на следващата таблица, не дават основание за отхвърляне на нито една от нулевите хипотези за равенство на средните стойности на индекса на трудност в отделните двойки тестови варианти.

Таблица 18. Резултати от дисперсионния анализ с повторни измервания на параметъра на трудност (*b*) в рамките на *IRT*

Двойка тестове	Тестов вариант	Сума от квадратите	Степени на свобода	Среден квадрат	Тестова статистика (*F*)	Статист. знач. (*p*)
1	вар. 92 вар. 128	0.053	1	0.053	0.555	0.458
2	вар. 96 вар. 132	0.114	1	0.114	1.230	0.270
3	вар. 146 вар. 110	0.011	1	0.011	0.0441	0.834

Налучкване на правилния отговор (c)

Резултатите от корелационния анализ на съответствията на стойностите на третия параметър на въпросите показват високи равнища на стабилност. Корелационните коефициенти при отделните двойки варират от 0.84 до 0.90

при нива на статистическа значимост p <0.05, както сочат данните от следващата таблица

Таблица 19. Коефициенти на стабилност (r_{xy}) на параметъра на налучкване на правилния отговор (*c*) в рамките на *IRT*

Двойка тестове	Тестов вариант	Коефициент на стабилност (r_{xy})	Статистическа значимост (*p*)
1	вар. 92 вар. 128	0.895	p < 0.05
2	вар. 96 вар. 132	0.874	p < 0.05
3	вар. 146 вар. 110	0.837	p < 0.05

Подобно на втората оценка на стабилността на дискриминативната сила, и тук резултатите от проверката на съотношенията между средните стойности на този параметър са в противоречие с очакваните. Данните в следващата таблица дават основание за последователно отхвърляне на нулевите хипотези за всяка двойка варианти на равнище p < 0.05. Това означава, че като цяло стойностите на параметъра за налучкване на правилния отговор при едната скала са изместени спрямо тези от другата със стъпка, равна на разликата между средните им стойности.

Таблица 20. Резултати от дисперсионния анализ с повторни измервания на параметъра на налучкване на правилния отговор (*c*) в рамките на *IRT*

Двойка тестове	Тестов вариант	Сума от квадратите	Степени на свобода	Среден квадрат	Тестова статистика (*F*)	Статист. знач. (*p*)
1	вар. 92 вар. 128	0.000	1	0.000	4.717	0.032
2	вар. 96 вар. 132	0.001	1	0.001	11.860	0.001
3	вар. 146 вар. 110	0.001	1	0.001	9.558	0.003

Допълнителна оценка на получените резултати ще бъде направена чрез извеждането на разликите между средните стойности на този параметър, на размера на ефекта и на статистическата мощност на критерия при всяка двойка тестови варианти.

При първата двойка (вар. 92 – 128) разликата между наблюдаваните

средни стойности е 0.002 и тази разлика води до размер на ефекта $\eta_p^2 = 0.046$, при мощност на критерия 0.576 (при $\alpha = 0.05$). С други думи, 4.6 % от дисперсията в стойностите на този параметър може да бъде обяснена с обстоятелството, че те са определени въз основа на различни извадки от и. л.

При втората двойка ((вар. 96 – 132) разликата между наблюдаваните средни стойности е малко по-висока (0.004), което води до по-голям размер на ефекта $\eta_p^2 = 0.108$, при мощност на критерия 0.926 (при $\alpha = 0.05$). Поради това 10.8 % от дисперсията в стойностите на този параметър може да бъде обяснена с обстоятелството, че те са определени въз основа на различни извадки от и. л.

При третата двойка (вар. 110 – 146) разликата между наблюдаваните средни стойности е 0.018, а размерът на ефекта е $\eta_p^2 = 0.088$, при мощност на критерия 0.865 (при $\alpha = 0.05$). С други думи, 8.8 % от дисперсията в стойностите на този параметър може да бъде обяснена с обстоятелството, че те са определени въз основа на различни извадки от и. л.

Ако бъдат приложени критериите на Дж. Коен, при първата двойка размерът на ефекта клони към среден/ умерен, при втората – по-скоро голям, а при третата – малко над среден.

3. Анализ на взаимовръзките между разноименните индекси/ параметри

Ще започнем анализа на допусканията за наличие на връзки между индексите, определени в рамките на *CTT*, и за отсъствие на такива връзки между параметрите, определени в рамките на *IRT*, чрез прилагане на подходящ за типа на съответната скала коефициент на корелация. При оценката на взаимовръзките между индексите D, p и r_{bis}, които участват в изследването със суровите си стойности, е приложен коефициентът на рангова корелация R на Спиърмън. При параметрите a, b и c е използван коефициентът на линейна корелация на Пиърсън. Резултатите, представени в следващата таблица, представляват съответните корелационни коефициенти, определени за всяка двойка статистики в рамките на отделните тестови варианти.

Общото впечатление от данните в таблицата, ако приложим класификацията на Дж. Хемфил, е, че равнището на взаимовръзките между отделните индекси е по-скоро високо.

Таблица 21. Взаимовръзки между разноименните индекси и параметри

Тестови варианти	Индекси по *CTT* (*R*)			Параметри по *IRT* (r_{xy})		
	D - p	***D - r_{bis}***	***p - r_{bis}***	***a - b***	***a - c***	***b -c***
Вар. 92	*0.324	*0.825	*0.405	*0.262	*-0.818	*-0.306
Вар. 128	*0.348	*0.815	*0.398	*0.302	*-0.781	*-0.273
Вар. 96	*0.402	*0.693	*0.637	*0.302	*-0.779	*-0.258
Вар. 132	*0.352	*0.739	*0.531	*0.231	*-0.713	*-0.240
Вар. 146	*0.512	*0.724	*0.554	*0.342	*-0.809	*-0.360
Вар. 110	*0.522	*0.808	*0.535	*0.445	*-0.787	*-0.311

Заб. Стойностите, маркирани със знака (*), са значими при p < 0.05

Над 85% от корелационните коефициенти имат стойности над 0.30, а някои от тях достигат до 0.83. Всички стойности са значими на ниво $p < 0.05$. Следва да обърнем особено внимание на значимите, високи равнища на взаимовръзка между параметрите на въпросите, определени в рамките на IRT.

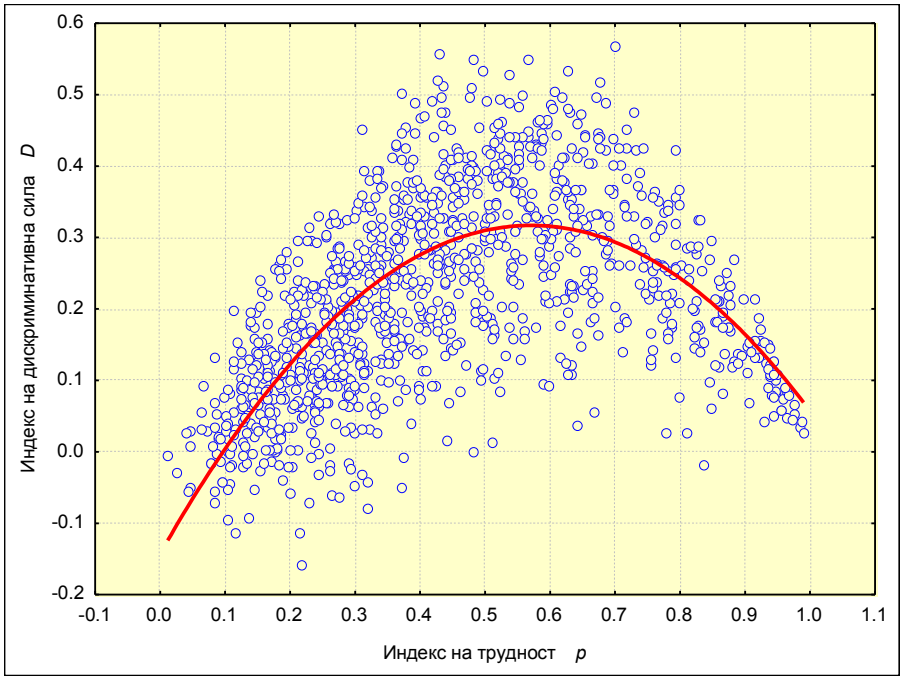
3.1. Взаимовръзки между индексите на въпросите в рамките на Класическата тестова теория

Но преди това нека да разгледаме взаимовръзките между индексите на въпросите в рамките на *CTT*. За отношението между индексите на дискриминативна сила D и трудност p беше изказано предположението за наличието на нелинейна връзка. За да се постави изследването на по-широка емпирична основа, беше направен повторен анализ на предполагаемата връзка чрез наблюдения върху трудността и дискриминативната сила на 1 200 въпроса, принадлежащи на 12 тестови варианта с номера 92, 96, 110, 127, 134, 141, 154, 166, 171, 175, 192 и 198. Стойностите на индексите са изчислени върху допускането, че въпросите от всеки вариант принадлежат към една скала (k = 100). На следващата графика е представена двумерна диаграмата на разсейването на въпросите.

Разсейването на въпросите дава ясна представа за това, че между двете характеристики на въпросите има връзка и че нейната форма е

нелинейна по своя характер. В зоната на екстремно висока трудност (p < 0.10), както и в тази с екстремно ниска трудност (p > 0.90), въпросите се характеризират с ниски абсолютни стойности на дискриминативния индекс. В съответствие с очакванията, въпросите в средната част на скалата на трудността, особено в интервала 0.40 – 0.70, се характеризират с високи стойности на дискриминативния индекс.

Фигура 3. Взаимовръзка между трудността *p* и дискриминативната сила *D*, определени по *CTT*



Прави впечатление, че отрицателни стойности на дискриминативната сила се наблюдават изключително в лявата част на хоризонталната ос (0.00 ≤ p ≤ 0.50), където са локализирани въпросите с по-висока трудност.

Негативни стойности на D в този интервал имат 51 въпроса, които съставляват 4.25% от всички наблюдения. Надясно от средата на тази ос въпроси с отрицателен индекс D (с едно изключение) липсват. Забелязва се струпване на по-голяма маса от въпроси в лявата част на хоризонталната ос, като въпросите с трудност 0.50 се намират на 64.75-тия процентил, а тези с трудност 0.60 – на 76.08-мия проценти. Подобно струпване има и в дясната част на скалата, в интервала 0.90 – 1.00, в който попадат 3.75% от въпросите.

За моделиране на взаимовръзката между променливите, формирани от

двата индекса, бе приложен метода на нелинейната регресия. В ролята на зависима променлива е дискриминативната сила (D), а на независима променлива (регресор) – трудността на въпросите (p). Трябва да отбележим, че зависимостта между двата индекса е, от една страна, безспорно нелинейна, тъй като промяната (нарастването) на p е свързано с непропорционална промяна в D. От друга страна, тя не следва да се разглежда като функционална, а като корелационна, тъй като в регресионния модел не участват всички фактори, въздействат на D, което води и до несъответствия между наблюдаваните и предсказаните стойности на този индекс.

За моделиране на връзката между индексите p и D бяха проверени няколко функции (претеглен метод на най - малките квадрати, Lowess и др.), от които бе избрана полиномна функция, зададена от следното регресионно уравнение:

$$D = (-0.143) + (1.612)p + (-1.411)p^2$$

Това е регресионен полином от втора степен, а оценката на параметрите е извършена по метода на най-малките квадрати. Изборът на тази функция за описване на връзката между двата индекса бе направен след проверка на полиномни функции с n = 2, 3, 4 и 5, а като критерий за избора бяха използвани статистическата значимости на оценките на параметрите и стандартните грешки на оценките $S_{y/x}$, които носят информация за големината на отклоненията на предсказаните от действителните стойности. При полиномните функции от трета и четвърта степен съответно един и два параметъра са статистически незначими (при този от трета степен $p(a_2)$ = 0.323, а при четвърта $p(a_0)$ = 0.955 и $p(a_1)$ = 0.824), а при полинома от пета степен нито един от параметрите не е значим при α = 0.05. При квадратичния полином се наблюдават най-ниски стандартни грешки на оценката, статистически значими стойности на оценките на всички параметри и най-тесни доверителни интервали, както е видно от следната таблица.

Описаната параболична форма на зависимост между двата основни индекса на трудност и дискриминативна сила позволява да се предсказват стойностите на D по наблюдаваните стойности на p. Основание за това дава и високата стойност на коефициента на корелация R, който отразява степента на

свързаност между двата индекса. Коефициентът R се определя като втори корен от R^2 и в това изследване той има стойност R = 0.686.

Таблица 22. Оценки на параметрите на квадратичния полином

Параметри	Оценка	Стандартна грешка	*t* (119)	p	Долна граница на довер. инт.	Горна граница на довер. инт.
a_0	-0.143	0.011	-12.529	0.000	-0.166	-0.121
a_1	1.612	0.052	30.844	0.000	1.509	1.714
a_2	-1.411	0.051	-27.461	0.000	-1.512	-1.310

Заб. Посочените граници са на 95% доверителен интервал при α = 0.05

Интересна е градацията на силата на различни типове корелационни коефициенти за съвкупността от 1 200 тестови въпроса, представени на горната графика. Докато Пиърсъновият коефициент на линейна корелация, е r_{xy} = 0.369, то стойността на коефициента на рангова корелация R е 0.478, а на корелационното отношение η за същите данни е 0.686. Това означава, че нелинейните модели са по-подходящи за описание на данните от линейния. Квадратът на нелинейния коефициент η^2 (ета на квадрат) е мярка за това каква част от дисперсията на зависимата променлива (условно това е дискриминативният индекс D) може да бъде обяснена чрез независимата променлива (условно индексът на трудност p). При анализираните данни η^2 = 0.471. Изразен в проценти, този индекс означава, че 47.10% от дисперсията на индекса D може да бъде обяснена чрез индекса p.

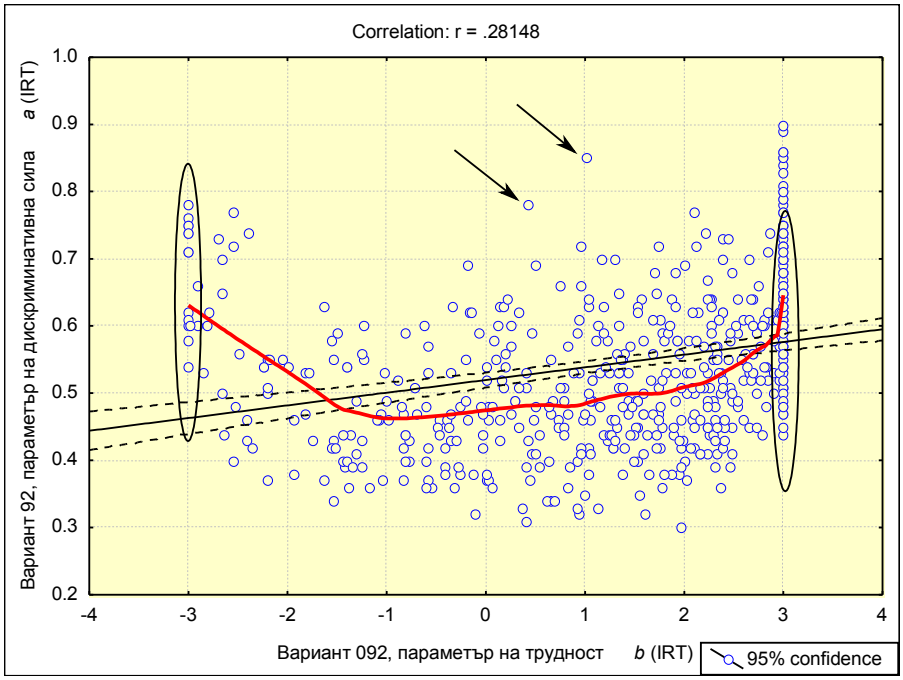
В случая, η^2 е по-скоро мярка за това доколко стойностите на p могат да се използват за прогнозиране на стойностите на D. Тъй като η^2 за нелинейните отношения има интерпретация, аналогична на тази на r^2 за линейните, разликата η^2 - r^2 би показала каква е степента на нелинейност в съвместното поведение на двете променливи (Калинов, 2010). По данните от вариант 92 степента на нелинейност η^2 - r^2 = 0.471 – 0.136 = 0.335.

3.2. Взаимовръзки между параметрите на въпросите в рамките на Теорията за отговор на тестов въпрос

Нека да се обърнем към взаимовръзките между параметрите на

въпросите, определени по *IRT*. Особен интерес представлява връзката между дискриминативната сила a трудността b, която може да бъде разгледана от една двойна перспектива. От една страна, това е очакването за липса на връзка между двата параметъра, а от друга – установената криволинейна връзка между съответните статистики в *CTT*. Корелационните коефициенти в таблица 21 говорят за наличието на умерено висока до силна (по скалата на Дж. Хемфил) взаимовръзка между параметрите b и a, като стойностите при отделните тестови варианти варират с от r = 0.23 (вар. 132) достигайки до r = 0.45 (вар. 110), всички значими при α = 0.05. Наблюдава се, следователно, ясно изразена позитивна, линейна взаимовръзка. Но дали характерът ѝ е чисто линеен, както бихме могли да предположим, водейки се от интервалния тип на скалата на двата параметъра? Анализът на диаграмите на разсейване на стойностите на двата параметъра поднасят поредната изненада. При всички анализирани тестови варианти се наблюдава, повече или по-малко, ясно изразен криволинеен модел на тази взаимовръзка. Ще илюстрираме нейния характер с диаграмата на разсейването, в която са агрегирани данните за a и b на шестте анализирани варианта, с общ брой от 598 валидни въпроса.

Фигура 4. Диаграма на разсейване на стойностите на *a* и *b*



На графиката се забелязва струпване на по-трудни въпроси в дясната

част на графиката (b > 0.00), характерно и за оценяването на тази характеристика чрез методите на *CTT*. Интересно е „сплескването" на облака от точки отдясно, което се изразява в подреждането на серия от въпроси в дясната част на графиката с трудност b = 3.00 (маркирани в овал). Подобен ефект се наблюдава и в лявата част на графиката, в която има по-малък брой въпроси с трудност b = -3.00 (също маркирани в овал). Това се дължи на особеност в алгоритъма на психометричния софтуер, който ограничава трудността до посочените гранични стойности. Забелязват се и две нетипични стойности (*outliers*), които имат такива координати, че отстраняването им би повишило слабо линейната корелация до от 0.231 до 0.248.

По-важна особеност на тази диаграма е контурът на облака от точки, който свидетелства за нелинейния характер на връзката между двата параметъра. Първоначално данните са апроксимирани по метода Lowess, а получената крива линия подсилва впечатлението от визуалния анализ. По-нататък върху данните бе приложен полиномен модел с n = 2, 3, 4 и 5, а като критерий за избора на подходяща функция бяха използвани статистическата значимости на оценките на параметрите и стандартните грешки на оценките $S_{y/x}$.

При апроксимирането с различни полиномни функции от втора, трета и пета степен бяха наблюдавани параметри, за които може да се предполага, че имат нулева стойност (p > 0.05). Единствено при функция от четвърта степен всички параметри са значими на ниво α = 0.05, с ниски стандартни грешки и тесни доверителни интервали. Данните от анализ са представени на следващата таблица.

Таблица 23. Оценки на параметрите на полиномната функция от 4-та степен

Параметри	Оценка	Стандартна грешка	*t* (119)	*p*	Долна граница на довер. инт.	Горна граница на довер. инт.
a_0	0.485	0.008	58.554	0.000	0.469	0.502
a_1	0.019	0.007	2.865	0.004	0.006	0.032
a_2	-0.012	0.005	-2.325	0.020	-0.022	-0.002
a_3	-0.003	0.001	-3.016	0.003	-0.005	-0.001
a_4	0.004	0.001	6.235	0.000	0.002	0.005

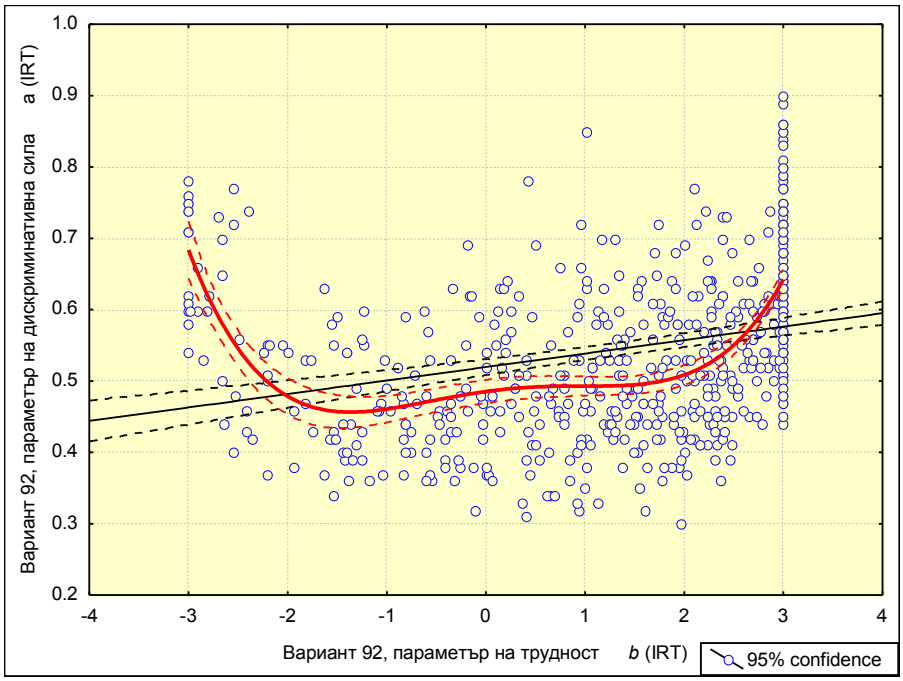
Заб. Посочените граници са на 95% доверителен интервал при α = 0.05

Съгласно данните от горната таблица, полиномната функция се задава от следното регресионно уравнение:

$$a = (0.485) + (0.019)b + (-0.012)b^2 + (-0.003)b^3 + (0.004)b^4$$

На следващата фигура е представена графиката на криволинейната полиномна функция, придружена от регресионната права на линейния модел на взаимовръзката между двата параметъра по данни от вариант 92. Дадени са и 95% доверителни интервали около двете линии.

Фигура 5. Взаимовръзка между трудността (*b*) и дискриминативната сила (*a*), определени по *IRT*

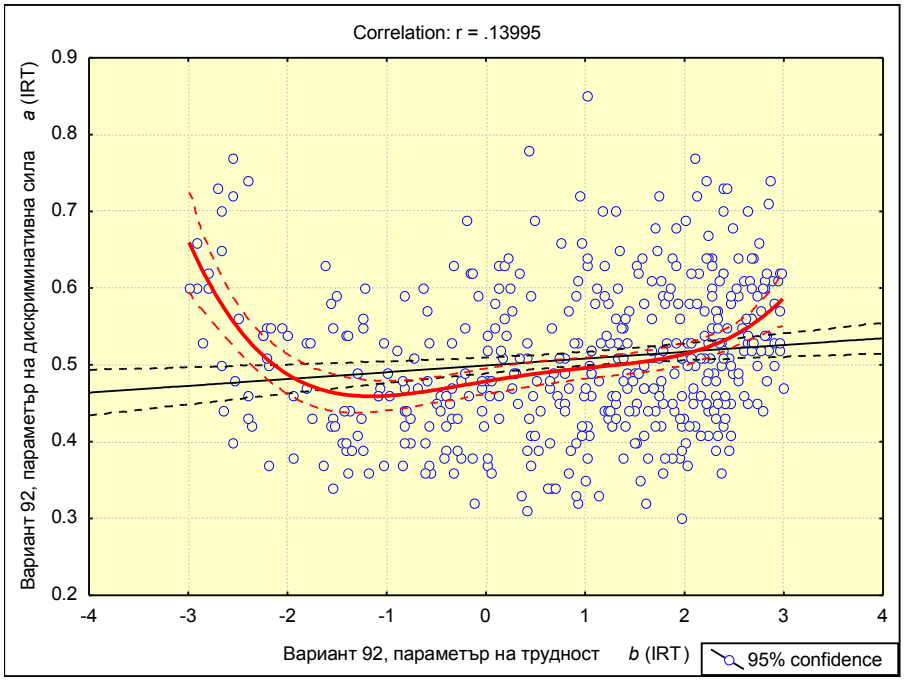


Ясно се забелязва тенденцията за рязко намаляване на стойностите на дискриминативния параметър (a) в интервала (-∞; -2.00), след което следва по-изгладена част в интервала (-2.00; +2.00), с тенденция към повишаване на стойностите на a и локален максимум в точка b = 0.00, следвана рязко повишаване на стойностите на дискриминативния параметър (a) дясната част на скалата, в интервала (2.00; +∞). Следователно, нелинейният характер на зависимостта между двата параметъра се проявява най-вече в двата края на

континуума, в зоните на високите (отрицателни или положителни) стойности на параметъра b. Частта от криволинейната графика в интервала (-2.00; +2.00), в която тя има по-добре изразен линеен характер, следва наклона на регресионната права, съответстващ на коефициент на линейна корелация r_{xy} = 0.281, изчислен върху данни от анализираните 598 валидни въпроса. Необходимо е да се отбележи, че този модел е в голяма степен устойчив и, с известни изменения, се наблюдава при всички изследвани тестови варианти.

По-горе обърнахме внимание върху наличието на множество въпроси с трудност b = -3.00 и b = 3.00. С такава екстремно ниска стойност са 13 въпроса (2.17% от анализираните 598 въпроса), а с екстремно висока стойност са 135 въпроса (22.50% от всички). Основателно е да се предположи, че „изкуственото" ограничаване на стойностите на b в тези граници би могло да доведе до изопачаване на силата и на действителната форма на взаимовръзката между двете променливи. Действително, след изключване от анализа на въпросите с посочените екстремни стойности на параметъра на трудност (b), т. е. с извадка от 450 въпроса, стойността на корелационния коефициент намалява от r_{xy} = 0.281 на r_{xy} = 0.140 при p < 0.05.

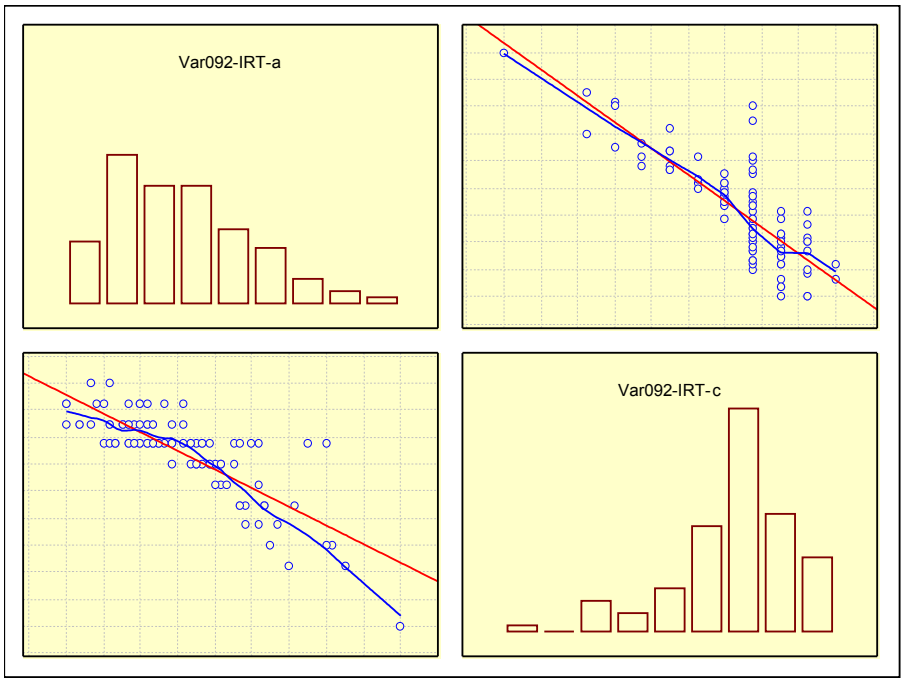
Фигура 6. Взаимовръзка между трудността (*b*) и дискриминативната сила (*a*), определени по *IRT*, след изваждане на въпросите с *b* = ± 3.00



Конфигурацията на точките обаче продължава да говори за нелинейна връзка между двата параметъра, както се вижда от горната графика, на която данните са апроксимирани с полиномна функция от четвърта степен.

Не по-малко интересни са получените данни за силни корелационни взаимовръзки между параметрите на дискриминативна сила a и налучкване на правилния отговор c. Корелационните коефициенти при различните тестови варианти варират от -0.713 при вариант 132 до -0.818 при вариант 92, при ниво на значимост $\alpha = 0.05$. Получените резултати говорят на негативна корелация между тези два параметъра – с увеличаване на вероятността от налучкване на правилния отговор дискриминативната сила намалява. Взаимовръзката между тези параметри може да се разглежда като линейна, макар че при всички диаграми на разсейване се наблюдава слабо изразена нелинейност, както е показано на следващата графика (долу вляво).

Фигура 7. Диаграми на разсейването и хистограми на параметрите на дискриминативна сила (*a*) и налучкване на правилния отговор (*c*) върху данни от вариант 92

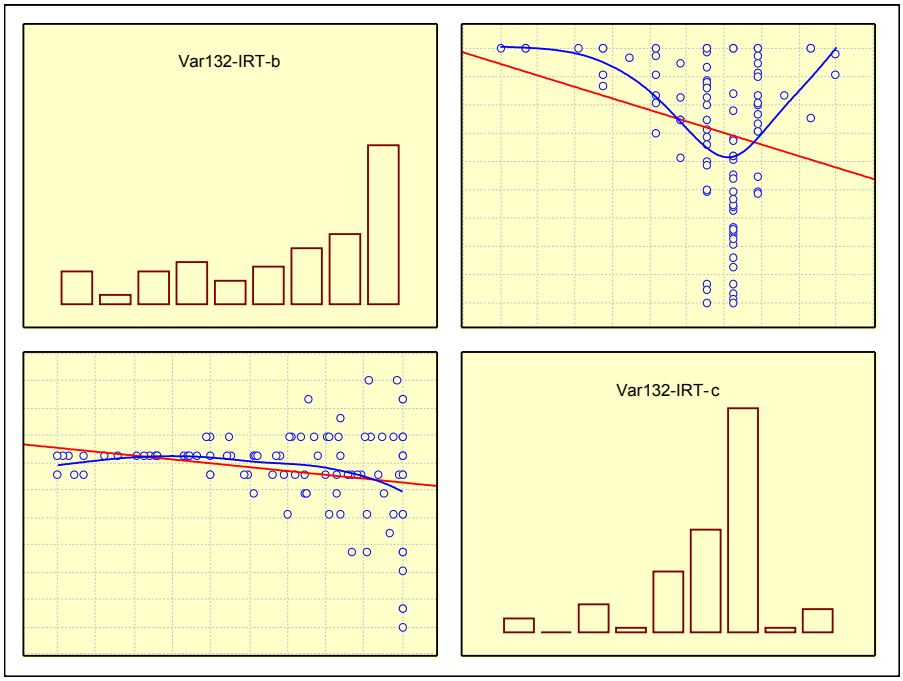


Интересно е, че тези особеност се наблюдава винаги и само тогава, когато условният „предиктор” е дискриминативната сила a, а „зависима променлива” е параметърът за налучкване c. При смяната на местата на тези

променливи на двете оси на диаграмата (горе вдясно) характерът на взаимовръзката е по-скоро линеен.

Наблюдават се и сравнително високи равнища на взаимовръзка между параметрите на трудност b и налучкване на правилния отговор c. Макар и по-слабо, отколкото с дискриминативния параметър, параметърът на налучкване c корелира с трудността на въпросите на равнища от -0.240 при вариант 132 до -0.360 при вариант 146, при ниво на значимост $\alpha = 0.05$. За взаимовръзката между тези два параметъра също е характерна обратнопропорционалната зависимост - с увеличаване на вероятността от налучкване на правилния отговор трудността намалява. Тук може да се прокара още един паралел с резултатите от предходния анализ. Диаграмите на разсейване на точките не са симетрични при смяна на местата на двата параметъра като условни „предиктори" и „зависими променливи". Променя се контурът на множеството от точки, а заедно с това и функцията, която може да се използва за апроксимация.

Фигура 8. Диаграми на разсейването и хистограми на параметрите на трудност (*b*) и налучкване на правилния отговор (*c*) върху данни от вариант 92



4. Анализ на съгласуваността между едноименните индекси и параметри

Може би най-важният аспект от съпоставителния анализ на очакваните характеристики на двете психометрични теории е проучването на степента на съгласуваност между функционално сходните им статистики. Такива са мерките за дискриминативност/ наклон и трудност/ позиция на тестовите въпроси. Като метод за изследване е използван корелационният анализ, но очевидно и тук не става дума за разглеждане на взаимовръзки от корелационен тип между едноименните статистики, а за оценка на това дали въпросите запазват своите позиции спрямо останалите в качеството им на индекси и параметри.

Ще започнем с общ преглед на коефициентите на линейна корелация, представени на следващата таблица.

Таблица 24. Коефициенти на линейна корелация между двойки едноименни статистики

	p - b	***D - a***	***r_{bis} - a***
Вар. 92	*-0.945	-0.161	-0.064
Вар. 128	*-0.970	-0.019	0.167
Вар. 96	*-0.970	*-0.294	0.012
Вар. 132	*-0.982	-0.099	*0.253
Вар. 146	*-0.985	*-0.424	0.011
Вар. 110	*-0.909	*-0.289	0.020

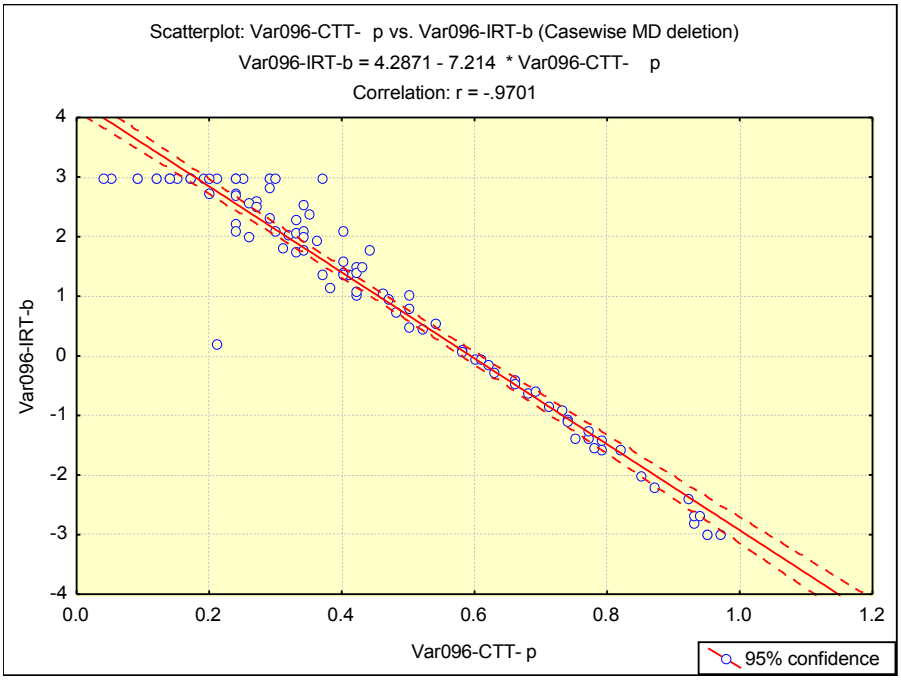
Заб. Стойностите, маркирани със знака (*), са значими при p < 0.05

На първо място прави впечатление контрастът между равнищата на стойностите, описващи съгласуваността между статистиките на трудността (p - b) и тези, описващи дискриминативната сила на въпросите. Корелационните коефициенти на съотношенията между p и b в различните тестови варианти са на равнища над 0.90, достигат до 0.97, като всички стойности са значими при ниво на α = 0.05. Очакван е отрицателният знак пред тях, както имаме предвид начина на изчисляване на трудността в рамките на *CTT*. Обратно, резултатите от съпоставянето на оценките на дискриминативната сила (D – a и r_{bis} - a) са неконсистентни и противоречиви. Корелационните коефициенти са сравнително ниски, при това малка част от тях са статистически значими, а при по-голямата част нулевата хипотеза не може да бъде отхвърлена.

Не по-малко интересни са формите на взаимовръзка между променливите. При анализа на диаграмите на разсейване на статистиките на

трудността (p - b) от различните субтестове беше установено, че формата на взаимовръзка може еднозначно да се определи като линейна, както е показано на следващата графика върху данни от вариант 96. На графиката се забелязва струпването на точки в интервала 0.20 - 0.40 по хоризонталната ос p, съответно в интервала 1.00 – 3.00 по вертикалната ос b, което свидетелства за преобладаващата трудност на въпросите в този вариант. Може да се забележи и отклонението от регресионната линия на няколко въпроса в интервала 0.90 – 1.00 по хоризонталната ос p, каквото се забелязва и при останалите тестови варианти. Като цяло обаче линейният характер на отношенията между двете статистики не буди съмнение.

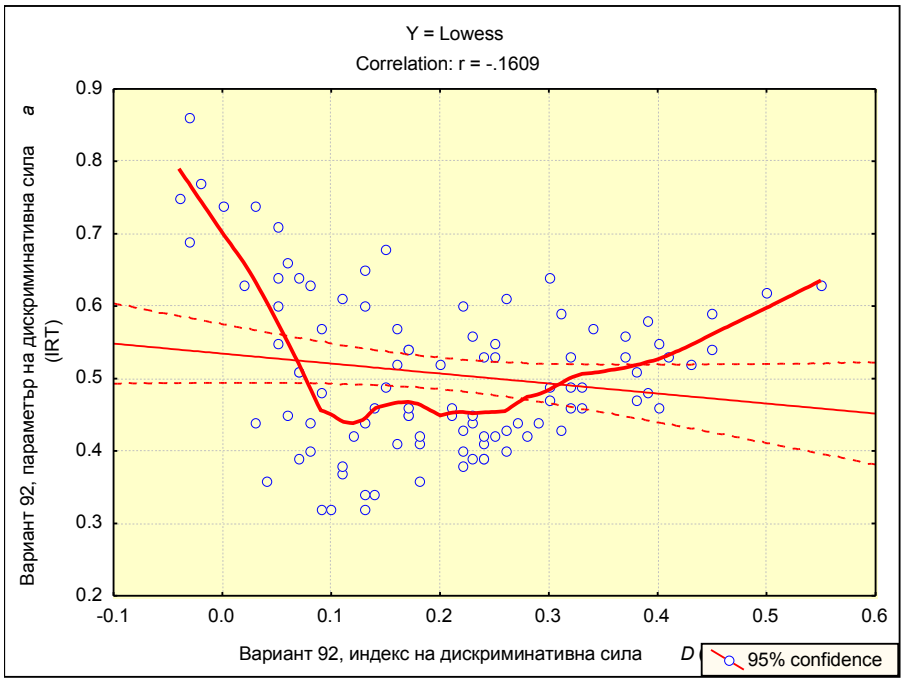
Фигура 9. Съгласуваност между статистиките на трудност *p* и *b* по данни от вариант 96



Не такава е картината при останалите двойки статистики на дискриминативната сила на въпросите (D – a *и* r_{bis} – a). Анализът на диаграмите на разсейването дава основание да се мисли, че сравнително ниските коефициенти на линейна корелация говорят не толкова за ниска степен на съгласуваност между техните стойности, а за нейния по-скоро нелинеен характер. На следващата графика нелинейния тип взаимовръзка е демонстриран върху данни за D по *CTT* и a по *IRT* от вариант 92.

За оценка на силата на нелинейната взаимовръзка между двете променливи беше използвано корелационното отношение ета (η). Неговата стойност по данните от вариант 92, представени на тази графика, е $\eta = 0.644$. Това означава, че между двете оценки на дискриминативната сила се наблюдава силна нелинейна корелация.

Фигура 10. Съгласуваност между статистиките на дискриминативна сила *D* и *a* по данни от вариант 92



Квадратът на нелинейния коефициент η^2 (ета на квадрат) е мярка за това каква част от дисперсията на зависимата променлива (условно това е дискриминативният параметър a може да бъде обяснена чрез независимата променлива (условно дискриминативният индекс D). При анализираните данни $\eta^2 = 0.414$. Изразен в проценти, този индекс означава, че 41.40% от дисперсията на параметъра a може да бъде обяснена чрез индекса D.

В случая, η^2 е по-скоро мярка за това доколко стойностите на D могат да се използват за прогнозиране на стойностите на a. Тъй като η^2 за нелинейните отношения има интерпретация, аналогична на тази на r^2 за линейните, разликата $\eta^2 - r^2$ би показала каква е степента на нелинейност в съвместното поведение на двете променливи. По данните от вариант 92 степента на нелинейност $\eta^2 - r^2 = 0.414 – 0.026 = 0.388$.

Впрочем, при диаграмите на разсейването на почти всички останали тестови варианти се наблюдава само един ясно изразен локален минимум на функцията Lowess. Поради това графиката на съвместното вариране на статистиките D и a може да бъде разгледана като съставена от две части, във всяка от които се наблюдава, както и на горната графика, сравнително ясно изразена линейна взаимовръзка. Наляво от локалния минимум корелацията е негативна, а надясно от него – позитивна.

Таблица 25. Коефициенти на линейна корелация между статистиките *D* и *a* наляво и надясно от локалните минимуми

Вариант	Локален минимум в т. …	r_{xy} наляво от лок. мин.	r_{xy} надясно от лок. мин.
092	D=0.12	*-0.718	*0.367
128	D=0.16	*-0.512	*0.518
096	D=0.26	*-0.506	*0.471
132	D=0.21	*-0.435	*0.558
146	D=0.24	*-0.591	*0.439
110	D=0.20	*-0.678	0.260

Заб. Стойностите, маркирани със знака (*), са значими при p < 0.05

Както се вижда от данните в горната таблица, минималните стойности на параметъра a при различните тестови варианти са в сравнително тесния диапазон от 0.12 до 0.26 от D. Корелационните коефициенти в последните две колони свидетелстват за наличието на висока степен на линейна съгласуваност както наляво, така и надясно от локалния минимум.

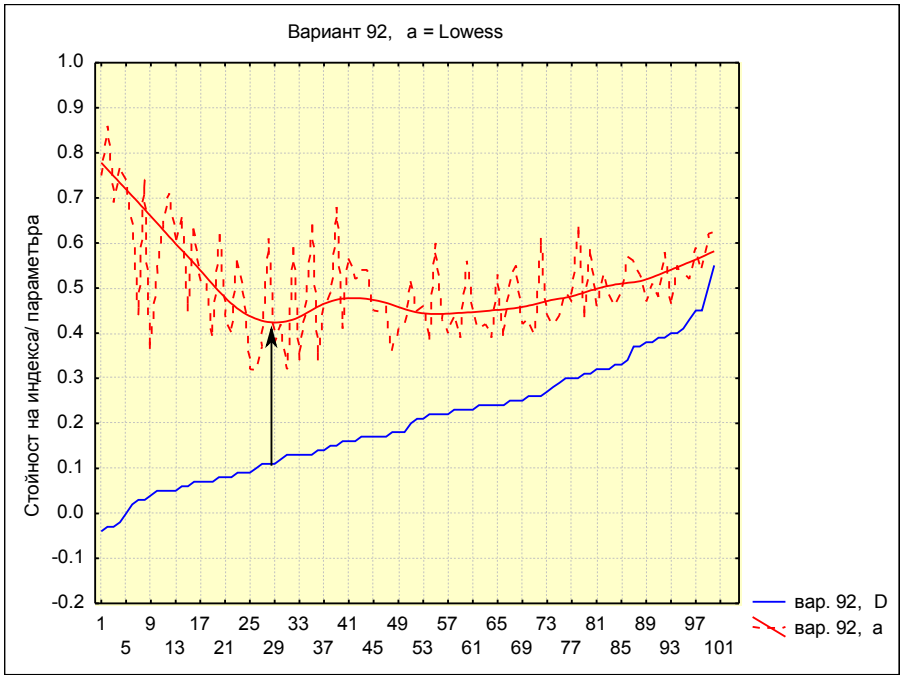
Нека да разгледаме следващата графика, на която са представени въпросите от вариант 92, подредени във възходящ ред по стойностите на индекса на дискриминативна сила D, и съответстващите им параметри a. Двете статистики са измерени в различни скали и съвместното им представяне има за цел единствено да визуализира изменението на параметъра a с нарастването на D.

Можем да забележим, че при по-голяма част от въпросите се наблюдава позитивна съгласуваност между двете статистики. Надясно от въпроса със стойност на D = 0.12, с нарастване на стойностите на D нарастват, макар и по-слабо, стойностите a.

При въпросите със стойности на D, по-ниски от 0.12, също се наблюдава

съгласуваност между двете статистики, но в този интервал тя е негативна. С намаляване на стойностите на D, стойностите a нарастват, дори с по-голяма интензивност. Както показват данните в горната таблица, съгласуваността между двете статистики наляво от съответния локален минимум е по-силно изразена, с по-високи, значими коефициенти на корелация, отколкото в интервала надясно от него.

Фигура 11. Стойности на индекса *D* и параметъра *a*, сортирани по възходящ ред на *D*



III. Дискусия

Анализът на данните от тестовите варианти доведе до множество интересни, в някои случаи изненадващи резултати, които не се съгласуват с направените предположения за очакваното поведението на статистиките на въпросите. Очертаха се и някои важни тенденции, които хвърлят нова светлина върху тяхното съвместно поведение.

Първият изследователски въпрос в анализа е свързан със стабилността, инвариантността на статистиките на въпросите, оценени в рамките на двете психометрични теории. За оценка на този аспект от тяхното поведение бяха

приложени два взаимно допълващи се подхода – корелационен анализ за изследване на относителната съгласуваност на стойностите на съответния индекс или параметър, получени в две различни условия, и дисперсионен анализ с повторни измервания – за оценка на съотношението между централните им тенденции при първото и второто измерване.

Най-напред следва да отбележим, че като цяло получените коефициенти на стабилност при всеки от наблюдаваните индекси и параметри, определени по двете теории, се характеризират с много високи равнища, които варират между 0.78 и 0.99. Всички емпирични стойности са статистически значими, разположени са в интервала над фиксирания праг от 0.70 и следователно могат да се разглеждат като свидетелство, че както индексите, определени в рамките на *CTT*, така и параметрите, определени в рамките на *IRT*, се отличават с висока степен на устойчивост, на стабилност по отношение на относителните им позиции при последователно оценяване в различни условия. Следва да се отбележи, че коефициентите на стабилност на всеки индекс или параметър, независимо върху коя двойка от тестови варианти са определени, се характеризират с близки, съпоставими стойности, което е още едно свидетелство за тяхната устойчивост и възпроизводимост.

Ако съпоставим равнищата на коефициентите на стабилност на различните статистики в рамките на всяка отделна теория, можем да установим няколко характерни особености.

В рамките на *CTT* системно по-високи са оценките на стабилността на индекса за трудност (p) на въпросите. При анализираните три двойки тестови варианти коефициентите надхвърлят 0.92, а при две от тях техните стойности са на равнища около 0.98. Сред тях е и най-високата наблюдавана стойност от 0.99, при варианти 96 – 132. Следователно трудността на въпросите може да се разглежда като тяхната най-устойчивата характеристика. По-малко устойчив е индексът на дискриминативна сила с коефициенти на корелация около 0.80 – 0.84, а най-малко – бисериалният коефициент с равнища на корелация около 0.78 - 0.81.

В рамките на *IRT* се наблюдава градация на коефициентите на стабилност, подобна на предходната. И тук със системно по-високи равнища се отличава стабилността на параметъра на трудност (b) на въпросите. При всички двойки тестови варианти коефициентите надхвърлят стойности от 0.92,

а при две от наблюденията достигат до 0.97. Малко по-ниски са равнищата при параметрите на дискриминативна сила (a) и на налучкване на правилния отговор (c), с коефициенти на корелация около 0.85 – 0.90.

Следователно може да се заключи, че двете тестови теории, приложени върху данните от ТОП, са в еднакво годни да осигурят висока степен на устойчивост на относителните равнища на статистиките на въпросите, оценени в различни условия.

Доста по-пъстри са резултатите, които произтичат от прилагането на втория критерий за устойчивост на статистиките на въпросите. При индексите, определени чрез алгоритмите на *CTT*, тестовите статистики и асоциираните с тях равнища на статистическа значимост не дават основания за отхвърляне на нулевите хипотези за равенство на медианите или средните стойности на съответните разпределения. Следователно, поставянето на тестовите въпроси в различни условия, при различни извадки от и. л., не предизвиква статистически значими разлики в средните равнища на техните индекси. Изключение от тази иначе добре подредена картина прави индексът на трудност (p) при една от двойките тестови варианти, при който се наблюдава значителен размер на ефекта при почти максимална мощност на критерия.

Изненада обаче поднася поведението на параметрите, определени в рамките на *IRT*. При два от тях – дискриминативна сила (a) и налучкване на правилните отговори (c), проверката на нулевите хипотези доведе до последователното им отхвърляне при всички двойки тестови варианти. Последвалата проверка за размерите на съответните ефекти, направена в съответствие с граничните стойности на Дж. Коен, дава основание те да бъдат оценени като средни или големи, при високи стойности на мощността на критерия, в повечето случаи надхвърляща 0.80. При тези параметри на въпросите може да се говори за отместване на наблюдаваните стойности на едното разпределение в сравнение с другото. Единственият параметър, който демонстрира стабилност, е този на трудността (b). Проверката на нулевите хипотези при трите двойки тестови варианти води до последователното им отхвърляне.

Може да се направи заключението, че отношение на втория критерий двете тестови теории, приложени върху данните от ТОП, не са равностойни. Прилагането на алгоритмите на *CTT* води по-често и при повече статистики на

въпросите до по-устойчиви резултати, отколкото тези на новата психометрична теория.

Ако съпоставим резултатите от анализите на едноименните статистики от двете психометрични теории, можем да обобщим, че най-устойчива статистика е трудността на въпросите. Макар че наблюдаваните коефициенти на корелация при отделните двойки варианти се различават, те се движат в един и същи интервал, непосредствено под максималната стойност от 1.00. Следователно, без значение коя тестова теория ще бъде приложена върху данните от ТОП, можем да очакваме най-ниска степен на вариативност при оценките на трудността на въпросите. Без да му придаваме по-голямо от необходимото значение, ще отбележим обстоятелството, че като цяло равнищата на стабилността на индексите по *CTT* са по-високи от тези на параметрите по *IRT*, както и това, че най-високата наблюдавана стойност от 0.992 е между две серии от индекси на трудността, определени по *CTT*.

Статистиките на дискриминативната сила са по-малко устойчиви от тези на трудността. В това отношение оценките на стабилността на параметрите по *IRT* са малко по-високи от тези на класическия дискриминативен индекс, който от своя страна има по-високи равнища от статистическия бисериален коефициент.

Ако разгледаме устойчивостта на статистиките на въпросите само от позицията на първия критерий, то можем да обобщим, че двете психометрични теории осигуряват в еднаква степен възпроизводимостта на относителните им позиции и в този смисъл са взаимнозаменяеми. Но ако приложим възприетия по-горе конюнктивен критерий за оценка, следва да посочим, че *CTT* „осигурява" стабилността на индексите на въпросите в 8 от наблюдаваните 9 случая (двойки тестови варианти), а *IRT* – в 3 от тези случаи. Следователно можем да направим заключението, че допускане 1 (а) за зависимост на индексите на трудност (p) и на дискриминативна сила (D, r_{bis}), определени в рамките на *CTT*, от извадките, въз основа на които са получени, не се потвърждава. Не намира опора в реалните данни и допускане 1 (б) за независимост на параметрите на дискриминативна сила (a), трудност (b) и налучкване на правилния отговор (c), определени в рамките на *IRT*, от извадките, въз основа на които са получени.

Но дали индексите на въпросите по *CTT* са действително независими от

извадките и какъв е психометричният смисъл на получените резултати? Нека да се обърнем към операционалните дефиниции на класическите индекси, както и към формулите за тяхното изчисляване. Те са пряко свързани с извадките, въз основа на които са определени техните стойности и поради това получените резултати хвърлят светлина и върху разпределението на съответната способност в извадките. Индексът на трудност (p) отразява относителния дял на правилните отговори на даден въпрос в извадката и неговата стабилност следва да се интерпретира като свидетелство, че този относителен дял се съхранява и възпроизвежда в различни тестови сесии, при различни групи от и. л. Дискриминативният индекс (D) е производен на предходния и отразява относителния дял на правилните отговори в "силната" и слабата" група. Малко по-ниската стабилност на този индекс отразява сравнително по-високата вариативност на относителните дялове в двете групи, в което определена роля може да има и налучкването на правилните отговори. Дискриминативният индекс отразява и друга особеност на извадката – нейната хомогенност по отношение на измервания признак. По-високи стойности на индекса се получават при хетерогенни извадки (поради различията в „силната" и „слабата" група), докато хомогенните извадки биха довели до понижаване на неговите стойности. Възпроизвеждането на равнищата на този индекс при различни извадки може да се обясни със съхраняването на приблизително една и съща структура на извадките по отношение на този признак. С други думи, устойчивостта на индексите е обусловена от относително стабилната структура на съответните компетентности на и. л., поне за определен период от време.

Очакваната стабилност на параметрите на въпросите по *IRT*, тяхната независимост от извадката бе подложена на съмнение, макар че авторите твърдят, както беше отбелязано по-горе, че стойностите на параметрите са характеристики на самите айтеми, но не и на групата, въз основа на която са изчислени. Нека да разгледаме този въпрос по-детайлно.

Характеристичната крива на въпросите се изгражда, подобно на трудността на въпросите по *CTT*, въз основа на относителния дял на правилните отговори, но не за цялата група, а за всяко равнище на наблюдаваната способност, т.е за всяка точка на скалата Θ. Теоретично, всяка

промяна в относителния дял на правилните отговори в различни точки от скалата (например, 5 правилни отговора повече в точка Θ = -2 и същевременно 5 правилни отговора по-малко в точка Θ = +2) би могла да доведе до промяна във формата на тази крива, т.е. в стойностите на параметрите, които я описват. Такава промяна обаче няма да се отрази върху стойността на класическия индекс (p). Следователно един от източниците на вариативност следва да се потърси в чувствителността на модела към промени в структурата на извадките.

В тази връзка следва да отбележим, че твърдението за инвариантния характер на параметрите произтича от допускането, че всички извадки са извлечени от една и съща популация (Harris, 1993; Embretson & Reise, 2000; Baker, 2001). Теоретично, тези извадки могат да бъдат и непредставителни и да бъдат извлечени от левия или десния край на популационното разпределение или от централната му част, т.е. да не споделят общи (еднакви) статистически характеристики. Тези извадки обаче са функционално еквивалентни, защото биха довели до едни и същи оценки на параметрите. Оттук може да се направи предположението, че наблюдаваната вариативност на параметрите се дължи на обстоятелството, че различните извадки са извлечени от различни генерални съвкупности. Анализът на поведението на индексите по *CTT* обаче подсказва, че такова предположение е по-скоро неоснователно. Следователно причините за наблюдаваната вариативност следва да се потърсят в особеностите на популационното разпределение. Както бе показано, анализът на неговата форма и размерност водят към заключението, че способностите не са разпределени нормално и не формират едномерно разпределение. А това са две от основните допускания, които формират фундамента на разглеждания модел на *IRT*. От друга страна, особеностите на реалните данните от ТОП не се съгласуват с тези допускания и поради това прилагането на този модел на *IRT* би било в противоречие с тях. Тук може да се добави и възможното влияние на обемите на извадки, както и адекватността на съответните характеристични криви за отделните въпроси.

Следва да се отбележи, че числовите стойности на параметрите, получени в хода на анализа на тестовите данни, не представляват действителните им стойности, а са техни оценки. Тяхната вариативност е

свидетелство за наличието на отклонения на наблюдаваните от действителните стойности. Тя не е основание да се подложат на съмнение теоретичните достойнства на изследвания модел на новата психометрична теория. Вариативността им обаче е свидетелство за негативния ефект от несъответствието между теоретичния модел и реалните данни.

В заключение можем да отбележим, че „меката" и по-„непретенциозна" Класическа теория се справя по-добре с проблема със стабилността на статистиките на въпросите, отколкото далеч по-сложната в концептуално, структурно и математическо отношение Теория за отговор на тестов въпрос. Отговорът се крие в характеристиките на реалните данни, които се характеризират с определена степен на неподреденост и неорганизираност, на аморфност, което се оказва по-значимо препятствие пред втория, а не пред първия теоретичен модел.

Един съпътстващ, но важен от методологична гледна точка въпрос беше вплетен в оценката на стабилността на статистиките – въпросът за типа на скалата, която образува индексът на трудността (p) по *CTT*, както и този на дискриминативна сила (D). Поради начина на изчисляване на техните стойности се приема, че тези скали са рангови. В съгласие с това схващане при анализите на стабилността на класическите индекси бяха приложени статистически методи, съответстващи на този тип измервания – коефициентът на рангова корелация R на Спиърмън и Знаково-ранговият тест на Уилкоксън за зависими извадки.

Успоредно с коефициентите на рангова корелация, за оценка на стабилността на класическите индекси бяха използвани и тези на линейна корелация. Общото впечатление е, че между оценките, направени по двата метода, се наблюдават незначителни разлики, като в почти всички случаи по-високи са равнищата на коефициентите на линейна корелация. Прилагането на тази мярка следователно не намалява, а подобрява оценката на силата на взаимовръзката между съответните две променливи.

Паралелно с числените методи, за определяне на вида на корелационните връзки, съответно на типа на скалите на индексите на трудност (p) и дискриминативна сила (D), бяха приложени и графични методи. Бяха изследвани диаграмите на разсейване на суровите стойности на двата индекса, получени въз основа на данните от вариантите в отделните двойки.

Анализът на диаграмите на разсейването също води към извода за интервалния характер на класическите индекси.

Към суровите стойности на класическите индекси бяха приложените два модела за апроксимация на данните – линеен и нелинеен, по метода на локално претеглената регресия. При двата индекса графичните методи, приложени паралелно, водят до почти едни и същи резултати.

При графичното представяне на криволинейните функции, описващи съвместното поведение на съответната двойка променливи, техните форми при двата индекса са изгладени, почти съвпадащи със съответните линейни регресионни прави. Само в отделни, тесни сегменти от тях взаимовръзката придобива монотонен характер, по-ярко изразен при дискриминативния индекс. Тъй като типът на скалата на индекса на трудност (p) беше изследван чрез отделна процедура за трансформация на суровите стойности, предположението за линейност/ интервалност на дискриминативния индекс (D) беше проверено допълнително чрез тестове за значимостта на регресионния коефициент, както и за цялостната годност на модела. Резултатите потвърждават предположението за линейния (интервален) характер на скалата на дискриминативния индекс. Впрочем, тези скали следва да се разглеждат като суб-интервални, съдържащи нелинейни сегменти, но приближаващи се до интервалния тип скали.

В заключение можем да обобщим, че скалите на индексите на трудност (p) и дискриминативна сила (D), определени в рамките на СТТ, не са интервални, но не са и рангови в строгия Стивънсов смисъл на тези понятия. В неговата класификация типовете скали са теоретични конструкти или “концептуални” скали по определението на М. Бунге (Бунге, 1975). Реалните скали обаче не съответстват на концептуалните дефиниции. Обсъжданите две скали очевидно се намират в „сивата” зона между ранговия и интервалния тип скали. При това тази зона може да бъде осветлена, т.е. да се направи конкретна оценка към кой скалов тип се приближава дадена скала. Би било по-обосновано тези теоретични конструкти (техните обеми) да се разглеждат не като обичайни множества, а като размити множества, съгласно концепцията на Л. Заде (Zadeh, 1965), според която членството на даден обект (скала) към дадено множество (скалов тип) е континуално, а не дихотомично.

Вторият изследователски въпрос в анализа е свързан с взаимовръзките

между разноименните индекси, от една страна, и параметри, от друга, определени въз основа на данните от отделни тестови варианти. Бяха направени две допускания за очакваното поведение на статистиките: 2 (а) за наличие на нелинейна взаимовръзка между стойностите на индексите на трудност (p) и на дискриминативна сила (D), определени в рамките на *CTT* върху една и съща извадка и 2 (б) за липса на взаимовръзки от корелационен или функционален тип между стойностите на параметрите на дискриминативна сила (a), трудност (b) и налучкване на правилния отговор (c), определени в рамките на *IRT*.

Първото допускане бе обосновано чрез модела „Данни единичен стимул" на К. Кумбс (Coombs, 1964). Второто допускане следва от вероятностния подход за оценяване на параметрите.

За оценка на взаимовръзките между индексите D, p и r_{bis}, които участват в изследването със суровите си стойности, бе приложен коефициентът на рангова корелация R на Спиърмън. При параметрите a, b и c е използван коефициентът на линейна корелация на Пиърсън.

Резултатите от направените анализи свидетелстват, че между всички статистики, определени в рамките на дадена теория, се наблюдават ясно изразени взаимовръзки от корелационен тип. Ако се приложи класификацията на Дж. Хемфил, общото равнище на силата на взаимовръзките между отделните индекси и параметри може да бъде оценено като високо. Преобладаващата част от стойностите са над горната граница от 0.30, а някои от тях достигат до 0.83, като всички стойности са значими на ниво $p < 0.05$. Следва да обърнем особено внимание на високите коефициенти на взаимовръзка между двете оценки на дискриминативната сила (D и r_{bis}) по *CTT*, както и между параметрите на дискриминативната сила a **и налучкване** c **по *IRT*, достигащи до стойности от 0.70 – 0.80**.

Особен интерес представляват взаимовръзките между индексите на въпросите в рамките на *CTT* и по-конкретно между индексите на дискриминативна сила D и трудност p, за които бе изказано предположението за наличието на нелинейна връзка. Оценките при отделните субтестове, направени чрез непараметричен коефициент на рангова корелация, са достатъчно високи (0.30+ – 0.50+), за да потвърдят това предположение – между двата индекса има поне някакъв тип монотонна връзка.

За да се разшири емпиричната основа на изследването, бе направен допълнителен анализ с данни за трудността и дискриминативната сила на 1 200 въпроса, принадлежащи на 12 тестови варианта. Интересна е градацията на равнищата на различни типове корелационни коефициенти, приложени върху тази съвкупност. Тази сила нараства от 0.369 при коефициента на линейна корелация на 0.478 при коефициента на рангова корелация, за да достигне до 0.686 при корелационното отношение η. Това означава, че нелинейните модели са по-подходящи за описание на данните от линейния.

Диаграмата на разсейването на техните стойности показва ясно, че двата индекса са свързани с криволинейна връзка, която има форма на изпъкнала парабола. Наблюдават се очакваните ниски абсолютни стойности на дискриминативния индекс в зоните на екстремно високи и ниски стойности на трудността, както и високи стойности на дискриминативния индекс в средната част на скалата на трудността.

Съвместното вариране на индексите беше апроксимирано по метода на нелинейната регресия с полиномна функция от втора степен. Беше определено и съответното регресионно уравнение с удовлетворителни оценки на параметрите. Степента на нелинейност в съвместното поведение на двете променливи (разликата η^2 - r^2), по данните от вариант 92, възлиза на 0.335.

Друга интересна характеристика, която се наблюдава в диаграмата на разсейването на двата индекса, е струпването на точки в лявата част на параболата, в посока към по-висока трудност и по-ниска дискриминативна сила на въпросите. Тук е разположена и основната маса от въпроси с отрицателни стойности на D. Този феномен заслужава по-внимателно обсъждане.

Трудността на въпросите отразява общия брой на правилните отговори на даден въпрос, следователно всяка точка, разположена по-наляво, отразява все по-намаляващия общ дял на правилните отговори. Дискриминативната сила отразява съотношението (разликата) между правилните отговори в двете екстремни групи, следователно всяка точка, разположена по-ниско, отразява все по-изравняващия се баланс между двете групи, който от даден момент нататък (D = 0.00) преминава в полза на „слабата" група. Например при въпрос 80 от вариант 192, който има най-ниската дискриминативна сила и е сред най-трудните в цялата съвкупност от 1 200 въпроса (D_{80} = -0.157, p = 0.216), делът

на правилните отговори в „слабата" група p_{low} = 0.283, а на тези от "силната група е много по-малък - p_{high} = 0.126. Как може да бъде обяснен фактът, че лицата с ниски способности се справят с трудните въпроси еднакво добре, дори по-добре, отколкото тези с високи способности? Отговорът може да бъде потърсен в третия, останал встрани от изследването индекс на налучкване на правилния отговор. В диаграмата на разсейването се съдържат нагледни свидетелства за това, че изпитваните прилагат стратегия за налучкване на правилния отговор, на която по-нататък ще обърнем допълнително внимание. Данните дават основание да смятаме, че тази стратегия е по-присъща на лицата от „слабата" група, отколкото на тези от „силната"

Подобно струпване на точки има и в дясната част на параболата, в посока към по-голяма лесната на въпросите. Безспорно и тук с намаляване на трудността на въпросите се наблюдава изравняване на съотношението между дела на правилните отговори в двете екстремни групи, изразяващо се в намаляване на стойностите на индекса на дискриминативна сила. Те обаче не преминават граничната нулева стойност, т .е. тази тенденция не се обръща.

Следователно може да се направи обобщението, че допускането за наличие на нелинейна връзка между индексите на трудността p и дискриминативна сила D, определени в рамките на *CTT*, намира опора в емпиричните данни.

Неочаквани се оказват резултатите от изследването на взаимовръзките между параметрите на тестовите въпроси, определени в рамките на *IRT*. Тук от особен интерес е взаимовръзката между дискриминативната сила a трудността b, която може да бъде разгледана не само от гледна точка на очакването за липса на връзка между двата параметъра, но и от тази на демонстрираната криволинейна връзка между аналогичните статистики в *CTT*. Получените корелационни коефициенти говорят за наличието на умерено високи до високи равнища на позитивна, статистически значима взаимовръзка между двата параметъра. При това направените оценки са свидетелство за равнищата на линейния компонент въз взаимовръзките при отделните тестови варианти. Анализът на диаграмите на разсейване показа, че по-подходящ е нелинейният модел на взаимовръзката, който може да бъде представен чрез полиномна функция от четвърта степен. Особеност на сложната крива е това, че нейният нелинеен характер се проявява най-вече в двата края на континуума, в зоните

на високите (отрицателни или положителни) стойности на параметъра b. В средната част на кривата, в зоната $\pm 2p$ кривата е почти изгладена, с локален максимум в точка p = 0.00, но с тенденция за повишаване на стойностите на a. Именно на тази част от кривата се дължи и линейният компонент във взаимовръзката.

Не по-малко интересна е наблюдаваната взаимовръзка между параметъра за налучкване на правилния отговор c и останалите два параметъра. При всички тестови варианти тя е негативна, статистически значима, с високи стойности, които при изследванията на взаимовръзката с дискриминативната сила a са системно по-високи, достигайки до равнища, надхвърлящи 0.80. При това, както и при взаимовръзката между a и b, получените оценки отразяват линейния компонент в съвместното поведение на тези два параметъра.

Диаграмите на разсейване на показват, че при двете комбинации от параметри (a – c и b – c) се наблюдава известна криволинейност на взаимовръзката, по-силно изразена при параметрите b и c, но, наблюдавана при различни тестови варианти, тя не се характеризира с определена устойчивост.

Следователно може да се направи заключението, че предположението за липса на взаимовръзка между параметрите на въпросите, определени чрез алгоритмите на *IRT*, не намира опора в емпиричните данни.

И така, параметърът на налучкване на правилния отговор c е негативно свързан с дискриминативната сила a и трудността b. С други думи, с увеличаване на стойността на този параметър се намаляват стойностите на останалите два. Нека да формулираме част от това твърдение по друг начин: с увеличаване на трудността на въпроса вероятността от случайно посочване на правилния отговор намалява.

Докато корелационните отношения са симетрични в статистически смисъл, т.е. $r\,(X,\ Y) = r\,(Y,\ X)$ (Калинов, 2010), то горните две твърдения не са симетрични, макар и да са еквивалентни. Смущаващо е не само това, че поведението на тези две статистики е съгласувано. Още по-смущаващо е това, че, съгласно резултатите, и. л. налучкват по-успешно коректните отговори на по-лесните въпроси, отколкото на по-трудните. Тази асиметричност може да

бъде преодоляна, ако във взаимовръзките между параметъра на налучкване (c) и другите два се допусне съществуването не на симетрична (корелационна), а на еднопосочна (каузална) връзка.

Включването на параметъра на налучкване в уравнението на характеристичната крива на въпроса от А. Бирнбаум (Birnbaum, 1968) отразява една житейска реалност – хората, които се явяват на изпит, могат да посочат правилния отговор, дори и когато нямат необходимите знания и умения. Чрез този параметър в зависимата променлива - вероятността за правилен отговор, се включва приносът на този феномен.

Съгласно трипараметричния модел на *IRT*, налучкването се разглежда като вероятност за посочване на коректния отговор по случаен начин. За даден въпрос този параметър е константен за всички точки на континуума на способностите Θ, т.е. лица с различни нива на способности имат равен шанс да посочат коректния отговор. От друг страна, принципът на инвариантност на параметрите, според който те са независими от извадката, е валиден и за параметъра c. По-точно, параметрите са независими от разпределението на и. л. на скалата на способностите Θ. Параметрите са атрибути на характеристичната крива на въпроса, които определят нейната форма и мястото на неговото функциониране на континуума Θ. Или, както беше отбелязано по-горе, параметрите характеризират въпроса, а не групата, отговорила на въпроса (Baker, 2001).

Дали обаче параметърът на налучкване c е, характеристика, която произтича само от въпроса? Действително, някои особености на въпросите като броят на дистракторите, начинът на формулиране на основата на въпроса и алтернативните отговори и др., влияят пряко върху този параметър. Но налучкването следва да се разглежда и като поведение, характеризиращо даден индивид или група индивиди. Ако застанем на тази позиция, връзката между параметъра на налучкване и другите два параметъра става по-прозрачна и може да получи правдоподобно обяснение.

Ако при отговора на даден въпрос определена група от лица прибегне до стратегията на налучкване, по-вероятно е това да не са всички лица от извадката, а само тези с по-ниски способности, които се намират на левия край на континуума Θ. Свидетелство за това, че до тази стратегия прибягват по-

скоро лицата от „слабата" група, отколкото на тези от „силната", беше намерено при анализа на взаимовръзката между индексите p и D по *CTT*. Резултат от прилагането на тази стратегия би бил „натиск" върху лявата част на характеристичната крива отдолу нагоре, който би повишил долната й граница. Повишаването на вероятността от правилен отговор в лявата част на скалата, без да е съпроводено с подобно повишаване в дясната й част, предполага промяна в разпределението на и. л. на скалата Θ, която се изразява в отслабване на дискриминативната сила на въпроса. Графично тази промяна се проявява чрез намаляване на наклона на характеристичната крива в точка Θ = b, трудността на въпроса. Това е механизмът, за който може да се предположи, че регулира съвместната вариация на двата параметъра.

Една от особеностите, свързани с въвеждането на третия параметър, е в промяната на начина на определяне на трудността, т. е. на позицията на въпроса на скалата Θ. Тя вече не е в точката, която съответства на вероятност от правилен отговор $P(\Theta) = 0.50$ (както при двупараметричния модел или при трипараметричния при $c = 0.00$), а в точка $P(\Theta) = (1 + c)/2$, т .е. средната стойност между стойността на параметъра c и неговия максимум 1.00. При въвеждане на третия параметър (или при промяна на стойността на c от 0.00 в по-висока стойност), позицията на въпроса би могла да се запази. Наличието на статистически значими негативни корелации между c и b обаче говори, че има тенденция въпросите да намаляват трудността си с увеличаване на вероятността от налучкване, т.е. да изместват позицията си наляво на скалата. Този феномен може да бъде обяснен с нарастването на общата маса на правилните отговори.

И така, може да се предположи, че силата, която отключва тази мрежа от взаимодействия, е налучкването на правилните отговори като съзнателно поведение на част от и. л. В англоезичната литература, освен термините"*guessing*" и "*pseudo-guessing*", се използва и терминът „*proneness to guessing*" в същото значение – склонност, предразположение към прилагане на такава стратегия, което може да характеризира единствено индивидите.

В този смисъл параметърът на налучкване може да бъде разглеждан, наред със способностите Θ, като втори параметър в *IRT*, който съдържа личностов компонент и който отразява определени психични процеси. Това

схващане се отличава от традиционното, съгласно което в рамките на *IRT* се разглеждат две различни, несвързани групи от параметри. Първата група включва само един параметър, който описва характеристиките на индивида (личностовия параметър Θ) и втора група от 1 до 5 параметъра, които описват ситуацията, в която е наблюдаван отговорът на този индивид (параметри на въпросите).

Такова поведение на индивидите може да бъде разгледано в светлината на теорията за мотивацията. Без съмнение, то е обусловено от мотивацията за постижения, която включва множество потребности и мотиви за действие, насочени към постигането на високи резултати и значими цели. Класическият модел на мотивационния процес представя избора на дадено поведение като последица от очакването, че това поведение ще доведе до определен резултат, и от субективната желателност или ценност на очаквания резултат (Величков, 1989; Дилова, 2008). Това е популярният модел "очаквания – ценност", представен от Х. Хекхаузен.

В изпитни ситуации като тази, при която са събрани анализираните данни, са налице и двата компонента. Кандидатстудентските изпити са с висок залог, имат състезателен характер и изходът от тяхното провеждане е дихотомичен. Положителният изход не се свежда просто до постигане на високо постижение (като конкретна изпитна оценка), а до преминаване в друг социален статус – този на студент. Това е очакваният резултат от изпита и без съмнение той е желан и има висока стойност за кандидатите. В този смисъл прибягването до стратегията за налучкване, чрез която даден кандидат може да придобие определено предимство и да надхвърли действителния си бал, може да се разглежда като мотивирано от очакването, това поведение би подпомогнало постигането на желания резултат. Разбира се, налучкването не е единственото действие, предприето от кандидатите за постигане на желаната цел. Тук е предмет на анализ, защото се реализира в тесните времеви рамки на изпита и намира пряко отражение в изпитните резултати.

И. Айзен и М. Фишбайн разработват по-детайлна теория на подбудителната регулация (за планираното поведение), която е по-подходяща като обяснителен модел (Ajzen, 1991, по Дилова, 2008). В тази теория, освен фактора „атитюд към поведението" (очаквания – ценност), въвеждат още два

фактора – „субективна норма” и „възприеман контрол”. Вторият фактор се отнася до субективно възприетата социална норма и готовността на индивида да съобрази своето конкретно поведение с нея.

Като цяло у нас социалната норма по-скоро толерира използването на, да кажем, непочтени методи за постигане на определено предимство. В една изпитна ситуация социалната норма се установява по-скоро от участниците в изпита (дори не само в конкретния изпит, а в изпитите като процедура за стратифициране), която също толерира използването на такива средства като преписване и подсказване, дори и при изпити, които имат явно съревнователен характер. Нормата за приемливо поведение може да бъде зададена и от конкретен авторитет – лицето (институцията), която провежда изпита. Този авторитет би могъл да ограничи прилагането на налучкване, например под страх от редуциране на наблюдавания тестов бал с определен „коефициент на налучкване”. Можем да си представим идеална изпитна ситуация, в която всички и. л., по силата на такава императивна инструкция, не прибягват до тази стратегия. В тази ситуация лицата, които нямат необходимата компетентност да отговорят на даден въпрос(в преобладаващата си част – в левия край на скалата Θ), няма да посочат отговор и ще получат нула точки. Поради това потенциалът на въпросите за налучкване, дори и да са двуалтернативни, няма да бъде реализиран. Ще бъде реализиран обаче двупараметричен модел с c = 0.00 при всички въпроси.

Обикновено изпитващите не поставят такива ограничения, дори напротив, при изпълнението на много тестови програми изпитваните са окуражавани да посочат какъвто и да е отговор на въпросите, по които не се чувстват уверени. Макар че при провеждане на Теста по общообразователна подготовка кандидат-студентите не са явно поощрявани да налучкват, няма ограничение, което да ги възпира. Още повече, че налучкването е най-„невинното” сред непочтените средства или поне най-малко ”видимо”.

Възприеманият контрол е оценката на индивида доколко е в неговите възможности да осъществи съответното поведение. Макар че на пръв поглед посочването по случаен начин на един от няколко алтернативни отговора не предполага ангажирането на големи или особени по характер поведенчески ресурси, в някои случаи такова поведение може и да не бъде реализирано. Не

са редки случаите, при които изпитваните не дават отговор на даден въпрос, на група въпроси или на части от теста, въпреки че биха могли да го направят.

И така, за преобладаващата част от кандидатите може да се предполага, че имат атитюд към поведението („желая да стана студент и ако налучквам, ще имам по-голям шанс да се класирам"), имат готовност за това социално приемливо поведение („всички в залата биха налучквали, когато са затруднени, а и това няма да бъде наказано") и са в състояние да реализират такова поведение. Мотивацията за високи постижения е свързана с една по-широка потребност на индивида от висока себеоценка, която да подсили положителния му Аз-образ. Подходящ обяснителен модел е социометричната теория, съгласно която себеоценката е показател за това как индивидът се чувства приет от околните (Leary et al, 1995, по Дилова, 2008). Има емпирични данни, че хората с по-висока себеоценка се чувстват по-добре приемани от другите, отколкото тези с ниска себеоценка (ibid.) Най-адекватният механизъм за поддържане и повишаване на висока себеоценка са действията, които носят добри резултати, т. е. високи постижения.

Тук възниква въпросът доколко удовлетворителни за индивида са високите резултати, постигнати (отчасти) с непочтени средства, например висок успех на изпита и влизане в университета, ако определен дял от бала се дължи на налучкване? М. Дилова говори за несъзнавани влияния на потребността от висока себеоценка върху познавателния процес, за определени „изкривявания" на себепознанието, които предпазват индивида от спадане на равнището на себеоценката и на Аз-образа (ibid., стр. 178). В редица експериментални когнитивни изследвания са получени свидетелства за това, че в някои случаи индивидите преработват информацията за себе си по начин, който им позволява да видят себе си в положителна светлина. Хората търсят положителната информация за себе си и избягват отрицателната, склонни са да виждат по-ясно положителните си страни и да ги надценяват, но не и отрицателните(ibid.)

Един подходящ обяснителен механизъм на желанието за налучкване в контекста на потребността от съхраняване и повишаване на себеоценката е феноменът на себеугодното атрибутиране. Той се изразява в стремежа за приписване на успехите на достойнствата собствената личност, а неуспехите- на външни фактори. При провал на изпит от класически тип (с преподавател,

застанал пред студента), последният би могъл да припише вината на преподавателя (много е строг, заяжда се, не ме харесва, задава трудни/ провокиращи въпроси и т. н.) или на късмета си (падна ми се въпрос, по който не бях подготвен, говорих след колега, който се представи блестящо и т.н.) При тестов изпит възможностите на студента да атрибутира евентуален неуспех екстернално са твърде ограничени. Един от достъпните начини за решаване на този вътрешен конфликтът е изпитваният да положи допълнителни усилия, тук и сега, по време на изпита, като използва възможността за налучкване, която самата изпитна форма му предоставя.

Третият изследователски въпрос в анализа е свързан със съгласуваността между съответстващите си индекси и параметри, определени в едно и също условие, т.е. при една и съща извадка от и. л.

Бяха направени допусканията, че между стойностите на оценките на трудността на въпросите p и b, както и между тяхната дискриминативна сила D и a и r_{bis} и a, няма съгласуваност. Това следва от теоретичната вариативност, нестабилност на индексите, определени в рамките на *CTT*, и тяхната инвариантност в рамките на *IRT*. Ако даден индекс варира в допустимите граници на неговото изменение, а съответният параметър е стабилен, не би могло да се очаква да има съгласуване между техните стойности.

Резултатите от направените корелационни анализи обаче водят към опровергаване на тези допускания. Наблюдаваните корелационни коефициенти, използвани като мярка за степента на съгласуваност между едноименните статистики, са свидетелство, че между стойностите, получени чрез алгоритмите на двете тестови теории, има определена степен на съответствие. Тук трябва да отбележим видимия контраст между степента на съгласуваност на статистиките на трудността и на дискриминативната сила на въпросите. Докато между оценките на трудността p и b се наблюдават, във всички изследвани тестови варианти, изключително високи, статистически значими коефициенти на линейна корелация, то при оценките на дискриминативната сила D и a и r_{bis} и a получените коефициенти на линейна корелация не са така еднопосочни.

Въз основа на получените оценки за степента на съгласуваност между p и b, на тяхната статистическа значимост, както и на анализа на диаграмите на разсейване на тези статистики можем да заключим, че между оценките на

трудността на въпросите, получени в рамките на двете тестови теории, се наблюдава висока степен на съгласуваност, която има ясно изразен линеен характер. С други думи, по отношение на трудността на въпросите Класическата тестова теория и Теорията за отговор на тестов въпрос могат да се разглеждат като взаимнозаменяеми.

Получените резултати подчертават на спецификата на трудността като базова характеристика на тестовите въпроси. Тук следва да обърнем внимание на особеното място, което психометричната общност отделя на еднопараметричния (Раш) модел на *IRT*, при който единственият вариативен параметър на въпросите е тяхната трудност (b). Няма съмнение, че при част от реалните ситуации (без значение дали преобладават или не) този модел не е адекватен на данните – въпросите се третират като притежаващи еднаква (фиксирана) дискриминативна сила (т. е. еднакъв наклон на характеристичната крива), игнорира се и проблемът с налучкването (параметърът c също е с фиксирана, нулева стойност). Той обаче притежава едно ценно качество – трудността е единственият параметър, който е разположен на скалата на способностите Θ. Нещо повече, въз основа на вероятностното моделиране на връзката между способността Θ и отговора на съответния въпрос, позицията на този въпрос (неговата трудност като единствена характеристика) може да бъде оценена независимо от това кои лица (т. е. какви индивидуални стойности на Θ) са използвани за тази оценка (Rasch, 2001). Самият Г. Раш, автор на този модел, радващ се на широка популярност, прави образно сравнение на този процес с измерването на температурата на даден обект, което трябва да води към приблизително едни и същи резултати, независимо от това какъв термометър е използван (ibid.)

Както беше отбелязано, различна е картината, която се очертава при изследването на съгласуваността на дискриминативните статистики. Получените оценки са неконсистентни и противоречиви, със сравнително ниски корелационни коефициенти, част от тях – негативни или статистически незначими. Но ако се обърнем към диаграмите на разсейване на съответните двойки статистики, тази неясна ситуация може да намери своето съдържателно обяснение. То се състои в това, че взаимовръзката между статистиките на дискриминативност при всички анализирани тестови варианти

има ясно изразен нелинеен характер, който не може да бъде експлициран, поне не толкова адекватно, чрез приложените линейни модели. Изразена чрез корелационното отношение η, при различните варианти степента на нелинейна съгласуваност между D и a е между 0.50 и 0.70, а степента на нелинейност (η^2 - r^2) варира от 0.30 до 0.50.

При по-детайлен анализ на диаграмите на разсейване на съответните две статистики на дискриминативната сила се очертават още една интересна особеност. При диаграмите на разсейване на почти всички тестови варианти се наблюдава само един ясно изразен локален минимум на апроксимиращата функция. Поради това графиката на съвместното вариране на статистиките на D и a може да бъде разгледана като съставена от две части, във всяка от които се наблюдава сравнително ясно изразена линейна взаимовръзка. В зоната наляво от локалния минимум на съответната функция корелацията е негативна, а надясно от него - позитивна. Интересно е да се отбележи, че като цяло минималните стойности на параметъра a са в областта около 0.20 от D, която се приема за долна граница на приемливост на стойностите на този индекс (Ebel, 1954). С други думи, негативната форма на съвместно вариране на D и a се наблюдава именно в зоната на неприемливите стойности на D.

Въз основа на резултатите от направените анализи на съвместното вариране може да се направи заключението, че между статистиките на дискриминативна се наблюдава определена степен на съгласуваност, макар и не така добре изразена, както при статистиките на трудността.

Цитирана литература

1. Abelson, R. P., Tukey, J. W. (1959). Efficient conversion of non-metric into metric information. *Proceedings of the Social Statistics Section of the American Statistical Association*. Washington, pp. 226-230.
2. Aiken, L. R. (1988). *Psychological Testing and Assessment*. Massachusetts: Allyn & Bacon, Inc.
3. Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50, pp. 179-211
4. Assessment Systems Corporation (1997). *User's Manual for XCALIBRE™ for Windows: Marginal Maximum-Likelihood Estimation Program*. St. Paul MN: Author.
5. Baker, B. O., Hardyck, C. D., Petrinovich, L. F. (1966). Weak measurements vs. strong statistics: An empirical critique of S. S. Stevens' prescriptions on statistics. *Educational and Psychological Measurement*, 26, pp. 291-309.
6. Baker, F. B. (2001). *The basics of Item response theory*. ERIC Clearinghouse on Assessment and Evaluation, 2-nd ed.
7. Birnbaum, A. (1968). Some latent trait models and their uses in inferring an examinee's ability. In: F. M. Lord and M. R. Novick (eds). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley, pp. 395-479.
8. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2-nd ed.). Hillsdale, NJ: Erlbaum.
9. Coombs, C. H. (1964). *A theory of data*. New York: John Wiley and Sons, Inc.
10. Ebel, R. L. (1954). Procedures for the analysis of classroom tests. *Educational and Psychological Measurement*, 14 (2), 352-364.
11. Embretson, S. E., Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
12. Fan, X. (1998). Item response theory and Classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, Vol. 58, No 3, pp. 357-381.
13. Frigg, R., Hartmann, S. "Models in science". *The Stanford Encyclopedia of Philosophy* (Winter 2006 Edition), Edward N. Zalta (ed.)

14. Gaito, J. (1960). Scale classification and statistics. *Psychological Review*, 67, pp. 277-278.
15. Gardner, P. L. (1975). Scales and statistics. *Review of Educational Research*, Vol. 45, No. 1, pp. 43-57.
16. Gribbons, B., Herman, J. (1997). True and quasi-experimental designs. *Practical Assessment, Research & Evaluation*, 5(14).
17. Hambleton, R. K., Jones, R. W. (1993). Comparison of Classical test theory and Item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12 (3), pp. 535-556.
18. Hambleton, R. K., Swaminathan, & H., Rogers, H. J. (1991). *Fundamentals of Item response theory*. Newbury Park, Ca.: Sage Publications, Inc.
19. Harris, D. (1993). Comparison of 1-, 2-, and 3-parameter IRT models. *Educational Measurement: Issues and Practice*, 12 (3), pp. 157-163.
20. Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist*, 58, pp. 78-79.
21. Leary, M. R. Tambor, E. S., Terdal, S. K. & Downs, D. L. (1995). Self-esteem as an interpersonal monitor: The sociometer hypothesis. *Journal of Personality and Social Psychology*, 68, pp. 518-530.
22. McNemar, Q. (1969). *Psychological statistics* (4th ed.). NY: John Wiley & Sons, Inc.
23. Rasch, G. (2001). On objectivity and specificity of the probabilistic basis for testing. In: *Rasch Lectures. In honor of Georg Rasch's 100 years birthday on the 21th of September, 2001*. Eds. L. Olsen and S. Kreiner. Copenhagen Business School.
24. Stevens, S. S. (1939). On the problem of scales for the measurement of psychological magnitudes. *Journal for Unified Science*, 1939, Vol. 9, pp. 94-99.
25. Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, Vol. 103, No. 2684, pp. 677-680.
26. Suppes, P. (1962). Models of data. In: E. Nagel, P. Suppes & A. Tarski (eds.). *Logic, methodology and philosophy of science: Proceedings of the 1960 International congress*. Stanford: Stanford University Press, pp. 252-261.
27. Wilkinson, L., & APA Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594-604.

28. Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, 8, pp. 338-353.
29. Анастази, А., Урбина, С. (2001). *Психологическое тестирование*. Санкт-Петербург: Питер.
30. Бунге, М. (1975). *Философия физики*. Москва: Прогресс.
31. Величков, А. (1989). *Личност и вътрешна мотивация*. С.: Издателство на БАН.
32. Дилова, М. (2008). *Експериментална психология на себепознанието*. С.: Нов български университет.
33. Калинов, К. (2010). *Статистически методи в поведенческите и социалните науки*. С., Нов български университет.
34. Стивенс, С. С. (1960). Математика, измерение и психофизика. В: С. С. Стивенс (ред.) *Экспериментальная психология*. Москва: Издательство иностранной литературы, стр.19-89.