Item response models in psychological assessment

Thesis Summary

„Al. I. Cuza” University

Faculty of Psychology and Educational Sciences

PhD Student: Cristian Opariuc-Dan

Coordinator: PhD Professor Constantin Ticu

Iași, 2013

A bad constructed test wrongly adjusted or invalid in psychology

is like a rusty scalpel in surgery:

even if it does not kill you it can leave permanent scars.

General structure of the thesis

The thesis has two parts, both including a practical section, followed by conclusion and discussion. **The first part**, covered in Chapters I-IV, illustrates the theoretical background of this thesis. We addressed issues related to general aspects of the psychological testing, presentation of item response models, how to build tests based on item response theory, including self-adaptive tests. **The second part** contains the results of BigFive Plus personality inventory adjustment for the item response models and comparisons between evaluation with conventional tests and evaluation by item response models.

Chapter I

General aspects related to psychological assessment. Historical perspective

We considered it necessary to start work with a short history of psychological testing in the first chapter, marking the main stages of the evolution of psychological tests (first experiments in psychology in the 19th century, the materialization of the "mental test" notion, the Alfred Binet moment, the emergence of non-verbal tests, collective tests as well as transition from abilities testing to personality assessment). Since the thesis aim was to compare the two theories, we recognized the value of an **overview of developments of the classical theory of psychological tests**, from the Spearman moment to the Gaussian distribution, Pearson's contributions and criticisms of Kelly. We ended our overview by describing the main postulates of classical test theory.

Item response theory was addressed in the same manner, in the next section**.** We illustrated the two major schools: the one initiated by Lord and Novick, the American school of thought and the European one, from Richardson and Rasch. The merger of the two was carried out by Professor Wright, contributing to the competitiveness of a number of contemporary researchers (Ace, Embretson, Reise, Hambleton, Van der Linden and many others). We did not fail to mention several Romanian researchers in this arena, among which Prof. Albu, Prof. Rusu, Prof. Pitariu, Prof. Balaszi, Prof. Dobrean from the Babeș-Bolyai University of Cluj-Napoca. The schools from Iasi and Timisoara are also represented by following Romanian researchers - Prof. Constantin, Prof. Havârneanu, Prof. Sava, Prof. Măricuțoiu and others.

In the last section we present the main distinction between item response theory and classical test theory, as they were mentioned by Susan Embretson and Paul Reise, adding our own views as well. Summing it up, in classical test theory, standard error of measurement is unique and applies to all scores, while in item response theory it is variable at each level of the continuum latent factor. In classical tests, the more items a classical test has, the more reliable it is; in response models we show that, a short test can be more reliable compared to the long ones. The classical theory claims that comparing scores is ideal if the forms are parallel. The item response theory states that comparing scores is ideal when coverage levels of the latent trait vary between individuals. To build a classic test we need representative samples. If we use item response models, analysis of items can be made without the use of representative sampling, even on simulated data. Classical tests grants significance only if the raw scores are compared with a norm. Item responses models do not require norms, meaning of the raw scores are given by comparing their distance to the items. Furthermore, the properties of the scale interval for conventional tests are achieved somewhat forced by the normal distribution. Response models acquire these properties by applying an appropriate measurement model. In the classical theory of test, mixed items determine an unbalanced total score, while the use of such items in the item response theory leads to an optimal model. Demonstrations of these differences are detailed in the paper and concluding that the item response theory is not an extension of classical test theory but a theory radically different from this.

We have also shown that in the response model the focus is not on the test, but on the item, the circular dependency issue found in conventional tests (subjects results are dependent on the item's sample and items properties are dependent on the sample of the subjects) being solved.

The first chapter ends with a summary of the main differences between the two theories, adapted from Hambleton and Jones.

Chapter II

Item response models

The second chapter aims to be an introduction to item response models, both theoretical and applied. The classical test theory is very simple, the measurement model being unique. Observed score is the sum of the actual score and measurement errors. Item response theory no longer provides the same simplicity, being a multi-model theory. The quality of the measurements depends heavily on the model chosen, as the one who best approximates the observed data, and the fulfillment of a number of assumptions. We began by describing the concept of latent trait, characteristics and significance of a measurement model and the item characteristic function. We felt it is vital to present item response theory assumptions: unidimensionality, local independence and the model of measurement. The applicative character of the chapter is given by the presentation of techniques for checking assumptions - Q_3 Yen test for local independence, for unidimensionality a series of heuristic techniques (scree-plot analysis and eigenvalue) and statistics (Stout test for essential unidimensionality, Martin-Löf, cluster and NOHARM methods). We detailed calculation formulas, most of the procedures described are implemented in the Psihosoft CATS system.

The paper continues to present **dichotomous and unidimensional models**, which are the most easily understood. Some models were presented: 1PL (Rasch) 2PL (Lord) and 3PL (Birnbaum), providing the item characteristic functions, curves, description and applicability. In order not to remain in the traditional approach we present other models of this type: ogival models, unused on practical applications but usefully to understands reasons of translating to logistic models, linear logistic model with latent trait (Fisher), four parameters logistic model with response time or model for repeated trials items. Although the number of item response models of this type is much larger, we did not continue the presentation because we have exceeded the estimated volume of the thesis.

Unidimensional polytomous item response models are the subject of a separate chapter. We defined the concepts of response categories and categorical intervals as well as the response function of the categorical interval and category response function. Even if they are more complex compared to the dichotomous, we were able to synthesize functions and characteristics curves of models such as: nominal response model (Bock), partial credit model (Masters), generalized partial credit model (Muraki), rating scale model (Andersen), graded response model (Samejima) and modified graded response model (Muraki). We have avoided, wherever possible, the use of sophisticated mathematical concepts and we synthesize the mathematical summary in terms of features, usability, and applicability.

Although item response theory stipulates unidimensionality, this assumption cannot perform every time. Therefore, there are **multi-dimensional models of the item response** with limited use and not sufficiently studied, but it may be used in the case of items that saturated more than one factor. First we distinguished between compensatory and non-compensatory models with partially compensatory variant of the latter. Then we treated multidimensional dichotomous models, namely multidimensional extensions of the 2PL and 3PL models, showing response surfaces of the items and their mathematical functions. A number of partially compensatory extensions of the dichotomous models were also discussed. In the case of polytomous models we presented multidimensional generalized partial credit model, multidimensional partial credit model and graded response multidimensional model. Towards the end, we mentioned other item response models, but without going into details.

The chapter ends by presenting some selection criteria in item response models, including a general decision-making scheme. Overall, we addressed a number of methods to study the adequacy of the model data; they will be detailed in the next chapter.

Chapter III

Construction of the tests based on item response theory

The third chapter is highly applicative and refers to the construction of tests based on item response theory. The section starts with the presentation of general and universal aspects concerning the **construction of psychological tests**. We have shown how to prepare constructs-map, defining constructs, map them, and how to operationalize. We then approached the elements of items design, presenting descriptive decisions and construct decisions as well as expert panel. A response space was then defined, the concepts of response or active pole and response or distracter pole, mentioning a number of techniques to develop space answers – phenomenography , SOLO taxonomy and Guttman scale.

In the last stage of design, **the choice of measurement model**, occupies the rest of this chapter. We showed the significance and properties of measurement scales in IRT, describing anchoring system, the logit scale, the scale in probabilistic unites and real scores scale. The conclusion is that measurement in the item response models differ radically from measurements using classical tests.

Item calibration aims to describe and intends providing practical guidelines on the main techniques for initial items calibration. Just for understanding concepts were presented a series of heuristics techniques, unused in applications, and then it continues with what are really important, methods based on maximum likelihood estimation. Parameter estimation techniques simultaneous for items and people (JMLE), maximum likelihood estimation method (MLE), marginal maximum likelihood method (MMLE) and Bayesian methods including an empirical distribution in the estimates were described in detail. Even if the mathematics is extremely complex, being able to support a thesis in the field, we tried to make it understandable by presenting and explaining relationships and by providing a concrete, clearly working algorithm. Thus, we wanted to empower the reader with a minimum knowledge of mathematics to understand and build their own tests based on item response theory.

Similarly we described the **methods for estimating the latent trait level of people**, basically the scoring system of item response theory. We detailed the easiest scoring method, maximum likelihood (ML), and two scoring systems used in professional applications, such as maximum a posteriori method (MAP) and expected a posteriori method (EAP), perhaps the most used in the present. We did not missed to include a non-iterative method of scoring - the method Owen - and the description of the role and the place the item and test information functions have in assessing the quality and accuracy of the assessment.

Chapter IV

Construction of the auto-adaptive tests

The next chapter considers applications from item response theory, especially through **self-adaptive tests**. The existence of numerous computerized tests on the market, some of questionable quality, is the reason for decision to mention some principles of construction of instruments for computer assisted psychological assessment. Thus, we have established the requirements of human-computer interface, detailing the system of the stimuli presentation, the system of responses and the requirements of the data management system. We also briefly presented the main evaluation methods using computerized tests built on item response theory - assessments with fixed and adaptive items.

Development of the items bank is a separate chapter because of its quality depends on the result of psychological assessment. We defined a number of characteristics of an item bank and its project, showing how to build a table of classification of items, how we can specify a set of constraints, how to calculate the objective function of the bank of items and how we know how many items needed at each level of latent factor to obtain an effective bank of items. We have also shown a number of methods to optimize the bank of items to obtain the maximum of the information function.

Initial and online calibrations are the next issues addressed. We showed that the items are not immutable; their parameters can be modified as a result of overexposure effect. This process, referred as deviation of parameters, can be monitored and attenuated by a number of techniques described in detail in the paper. Perhaps one of the curiosities of self-adaptive tests is the way in which items are selected and how they adapt to the subject's answers. This curiosity will be satisfied in the section for automatic selection of items. We have shown some strategies to entry a test, a series of methods and techniques for selecting the next item, the advantages and disadvantages of each, and some methods to complete the assessment. There have been also given a number of techniques to control exposure and balancing the items in order to reduce obsolescence of the bank of items, as well as methods for identifying aberrant response pattern, item response theory having strong mathematical mechanisms to control the facade trends or random responses.

The role of the first four chapters was to create the conceptual, theoretical, base for the construction of tests based on item response theory and provide a set of practical methods to achieve this. This was achieved in a total of about 190 pages; important aspects are dealt with in detail. Some elements (such as response models important for psychology or certain techniques) were briefly covered and perhaps deserved more attention. The latest researches published in recent issues of the journal Psychometrika were not included in the thesis. Reasons for this were their abstract nature and the thesis's word limit. There was the risk of presenting a too voluminous paper, containing elements aimed at few specialists in this field, without a general population interest.

Chapter V

Influence of psychological assessment model on accuracy and reliability of results

Chapter five contains over 200 pages and deals strictly with a practical adaptation of a personality inventory (BigFive Plus) to item response theory and the study of relations between classical assessment and assessment based on item response theory. Originally, we wanted the analysis of two instruments: BigFive Plus personality inventory and EVIQ intelligence test. We chose not to analyze EVIQ for the following reasons. First, the analysis would have doubled the volume of chapter. Second, most research on item response theory was performed using aptitude tests, creating the false impression that the item response models can be applied only to those instruments. We have extended the scope, proposing the term "coverage level of the latent trait" instead of difficulty and showed that item response theory can be used without problems in the case of personality tests as well. EVIQ implementation in Psihosoft CATS will be carried out separately, as a result of further study.

The overall objective of this thesis is to investigate the compatibility of evaluations based on item response models and those based on the classical theory. In order to achieve this general objective we follow several steps. First, we analyze the test to study the assumptions of item response models; second, we build a computerized evaluation system based on item response theory, third, we see the degree of compatibility between the scores on the items of the classic test and IRT and we will see whether we can speak of a relationship between the psychometric properties of classical and IRT tests.

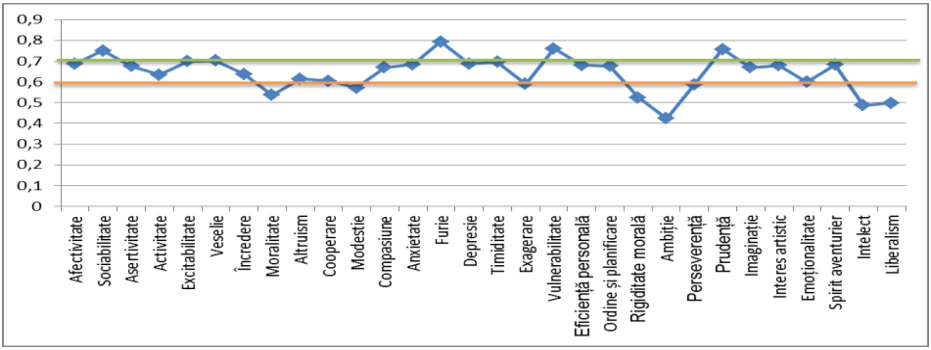
Research design involves two studies. In the first study we use classical techniques for construction of psychological tests and items calibration. Thus, the 240 personality items of the inventory will be analyzed at the item and scale level, studying the normal distribution, internal consistency and factorial structure. Then we proceed to analyze unidimensionality and initial calibration using item response models type 2PL or 3PL. Ideal would be to use a 3PL model type, but this was not possible every time. The second study involves the administration of the classical test and those based on item response theory to the same group of subjects, at a certain period of time, and study the relationship between scores, the parameters of discrimination and the level of coverage in latent trait.

Research hypotheses are simple, clear and precise. These are in line with studies (relatively few) on this subject. We note Lawson’s (1991) and Xitao’s (1998) research that uses ability tests and linear relationships. Embretson and Hambleton latest research was conducted (2009 and 2011) on simulated data predicting the type of relationship results in our studies. The null hypothesis states that there is no relationship between the results of classical tests and the administration of IRT tests. Rejecting the null hypothesis would lead to the presence of relations on two levels: the scores - there are significant links between the scores achieved by the classical version and the IRT version - and at the items parameters - discrimination and coverage in latent trait. That's why we don’t use a Rasch model measurement type. Discrimination parameter could not be studied. Methods of analysis in the case of the second study are not sophisticated. We investigate the linear nature of the relationship by Bravais-Pearson r bivariate correlations and the existence of differences by Student t test for paired samples. Since we can assume that the relationship can exist without having a linear character, we shall proceed to linear regression of variables to one another through processes such as estimating the curve. The principle of minimal residues will indicate the best relationship. Data analysis programs used are IBM SPSS for Windows and Psihosoft CATS, the latter being used in tasks that require especially item response theory techniques.

Research samples are different. The first study were used a number of 4647 subjects, and for the second study group of 323 students, their characteristics are described in each study.

Analysis of the normality of the distributions was performed at the level of all the 30 faces and to the five factors of the instrument. Tests were used to compare the observed distribution with the theoretical normal distribution (Kolmogorov-Smirnov), analysis of symmetry and excess coefficient and distance analysis of the observed data with the regression line in relation to the normal distribution. The results are presented in detail in the paper and lead to distributions that deviate significantly from the normal distribution.

Scale consistency analyses were performed similar to univariate normality, both on dimensions and factors. We also studied the presence of multiplicative interactions using Tukey test and multivariate normality of distributions using Hotelling t^2 test. It is noted that a relatively small number of factors are above the threshold of .7 required for a consistent scale. Most factors have an internal consistency between .6 and .7, which is acceptable for research purposes, but questionable for diagnostic. Also, a number of 8 factors have small levels of consistency; normally they should be excluded from the analysis. Multivariate normal distribution criterion was reached for all variables analyzed, indicating the relevance of the method. Items are non-additive, but multiplicative; however this is not an error but is caused by the dichotomous nature and the small number of items in the facets.



We have not been limited only to analysis of consistency, but we proceeded to investigate the internal structure of the 30 factors as well. Since classical factor analysis cannot be used in optimal conditions due to lack of normality, the presence of multiplicative interactions and low consistency, we used a nonparametric method based on vector and centroide coordinates analysis. This method is called **categorical principal components analysis (CATPCA)** and described in detail in the article "Principal components analysis for categorical data" in the Journal Psychology of Human Resources, Volume 10, no. 2/2012, pages 103-117. This study deals with a significant amount of data and is accompanied by a critical analysis of items for each factor. The result was a number of three factors that will be completely excluded, 16 items to be removed, and only 3 factors having a purely one-dimensional nature. Most factors present a dimensional-axial structure or two-dimensional structure. The presence of an axis means that the second dimension has not the specific of a component, but orientates the main dimension. Axis-dimension distinction was made following items saturation analysis, investigating centroids coordinates and critical analysis of clusters of items. The results of this analysis will be published in the journal "Annals of the Alexandru Ioan Cuza University" series Psychology; article is currently in the process of reviewing.

We suggested hypothetically the exclusion of items or factors. After studying the internal structure of the factors, **dimensionality analysis and calibration of items** followed. Unidimensionality was verified by DIMTEST, in partitioning set were included the items that strongest saturates the factor, other items were included in the evaluation set. Following this analysis, the problematic items were effectively removed, unidimensionality controls being made by NOHARM. The results support strongly the CATPCA analysis. Indeed, three factors have been completely eliminated, most losing one or two items in order to reach a definite one-dimensional structure. Calibration considered the 3PL model, the assumption of the model being tested by measuring the ratio of likelihood logistics. In the case of some latent traits, calibration failed for the 3PL model and we use the Lord (2PL) model. Unfortunately, a single latent trait strictly complies with the requirements of 2PL model, morality factor. For all other factors, the distribution of observed data at the item level deviates significantly from the model characteristic curve. Items also showed a tendency to concentrate in middle area of the latent trait continuum for every factor. Both biases are the result of the origin of the items from classic tests and hold both construction mode of the instrument and data collection. Even if the results clearly indicate the presence of errors, they have led, however, to useful results.

The second study aims to verify the following hypothesis. **The first hypothesis** supports a link between the latent factor level of the subject assessed with classical test and the subjects evaluated by IRT. We note that classic test was administered to all the 240 items, paper and pencil format and computerized test has a smaller number of items, missing three factors, and the items were presented randomly. Between the two administrations, there was a period of 4-5 months. Variables were the z score of each subject at each factor and the estimated theta for each factor. The comparison is possible because the distributions are standardized and is strongly compatible. Analyses were represented by descriptive techniques, differential and regressive. For item response models, the estimators averages focuses on the middle of the latent trait continuum, showing the origin of the items. Standard errors are very small, as well as the standard deviations. The amplitudes of the distributions are consistent with this orientation on the average of the latent trait. For the classical items the amplitudes of distributions are much larger, sample dependence is obvious. Assessing subjects with an IRT test, we could conclude average levels of latent factor, without emphasis on most people. Using a classic test and a norm built on the 323 subjects, some people would present very high or very low levels of latent factors, and in reality this is wrong. Significant differences have resulted between the results obtained with classical tests and IRT tests in all latent factors, which supports once again the dependence of the sample. There is, however, a number of significant linear correlations, only 4 factors showing that there is no significant relationship between variables. Nevertheless, the best explanatory model is not linear but cubic and logistic models. Cubic models are characteristics of a third degree equation, and those logistic correspond to an inverse of an exponential equation. In our research, along with the nature of the relationship between scores obtained in tests built on two theories, we have provided the equations of transformation of scores based on cubic and logistic models. These results do not invalidate the research conducted by the authors mentioned but complement them, claiming cubic models resulting from simulated studies. Despite the biases present, we could argue that from an assessment using a classical test and evaluation with a variant of IRT, the results are consistent even on a linear relationship, but the best model is not linear but a cubic or logistic depending on the nature of the latent factor measured.

Tabel V-198 Relații între nivelurile factorului latent pentru probe IRT și cotele z ale scorurilor brute pentru probele clasice

FACTOR LATENT	R²	F	DF1	DF2	SIG.	CONST.	B1	B2	B3	MODEL
AFECTIVITATE	0,508	109,799	3	319	0,000	-0,035	0,125	-0,015	-0,010	Cubic
SOCIABILITATE	0,804	436,977	3	319	0,000	0,201	-0,217	0,014	0,017	Cubic
ASERTIVITATE	0,734	293,811	3	319	0,000	0,231	0,218	-0,035	-0,002	Cubic
ACTIVITATE	0,167	21,309	3	319	0,000	-1,724	0,796	0,140	-0,228	Cubic
EXCITABILITATE	0,730	286,910	3	319	0,000	-0299	-0,567	0,081	0,020	Cubic
VESELIE	0,565	138,140	3	319	0,000	-0,840	-0,499	0,106	0,098	Cubic
ÎNCREDERE	0,747	312,923	3	318	0,000	-0,015	0,052	0,002	0,001	Cubic
MORALITATE	0,665	382,701	1	193	0,000	6,037	3,683	-	-	Logistic
ALTRUISM	0,789	397,370	3	318	0,000	-0,340	0,166	0,015	-0,001	Cubic
COOPERARE	0,333	159,558	1	320	0,000	1,059	1,517	-	-	Logistic
MODESTIE	0,168	21,456	3	318	0,000	-1,244	-0,282	0,348	0,066	Cubic
COMPASIUNE	0,021	6,802	1	320	0,010	7,542	1,052	-	-	Logistic
ANXIETATE	0,234	98,220	1	321	0,000	0,744	1,469	-	-	Logistic
FURIE	0,703	252,017	3	319	0,000	-0,005	0,003	0,000	0,000	Cubic
DEPRESIE	0,294	44,185	3	319	0,000	0,259	-0,090	-0,008	0,005	Cubic
TIMIDITATE	0,653	200,109	3	319	0,000	-0,298	0,068	0,004	0,000	Cubic
VULNERABILITATE	-	-	-	-	-	-	-	-	-	-
EFICIENȚĂ	0,418	76,249	3	319	0,000	-0,088	0,240	0,026	-0,005	Cubic
ORDINE	0,623	175,736	3	319	0,000	0,217	-0,057	-0,001	0,003	Cubic
PERSEVERENȚĂ	0,526	118,028	3	319	0,000	-0,257	0,325	0,023	-0,013	Cubic
PRUDENȚĂ	0,561	135,844	3	319	0,000	-0,139	-0,018	0,002	0,000	Cubic
IMAGINAȚIE	0,701	249,125	3	319	0,000	-0,277	0,081	0,004	-0,002	Cubic
INTERES ARTISTIC	0,110	13,187	3	319	0,000	1,602	0,230	-0,433	0,042	Cubic
EMOȚIONALITATE	0,060	6,811	3	319	0,000	-0,109	-0,008	0,032	0,011	Cubic
AVENTURĂ	-	-	-	-	-	-	-	-	-	-
INTELECT	0,204	27,235	3	319	0,000	-0,538	-0,275	0,065	0,026	Cubic
LIBERALISM	0,281	41,607	3	319	0,000	0,033	0,263	0,029	-0,016	Cubic

The second level of analysis focused on the psychometric properties of the items, the coverage level in latent trait and discrimination. **The second hypothesis** states that there are differences between discrimination of classic items and items built on item response theory. Discrimination of the classic items can be evaluated based on a point-biserial correlation between item and scale. However, this indicator cannot be compared directly with the discrimination parameter for item response models because of different scales. Therefore, the common denominator is the logistic scale, Fisher transformed of the point-biserial correlation coefficient bringing data to a common denominator. The analysis was performed at each dimension and for the entire test, using the same techniques. The amplitude of discrimination parameter for classic items is much lower compared to that of items IRT, the discrimination mean of the second category being, also, upper in terms of very low standard errors of the estimates. The same parameters fall and standard deviation, elements that lead us to the idea of a superior discriminative capacity of IRT items compared to the classics. The relationship between these variables also has a linear character, but the best explanatory model is the cubic corresponding to an equation of the third degree. This model is also preserved at the dimensions, not only to the whole test.

Clearly, significant differences exist between the two parameters, the tests having different capabilities of discrimination, higher for item response models.

Tabel V-210 Medii și abateri standard în cazul întregului instrument

		Fisher (rp.bis) N=323	Discriminare item (a) N=4043
N	Valid	188	188
	Absent	0	0
Media		,64295	1,55179
Eroare standard medie		,009302	,048290
Abaterea standard		,127549	,662116
Minimum		,216	,172
Maximum		1,013	5,708

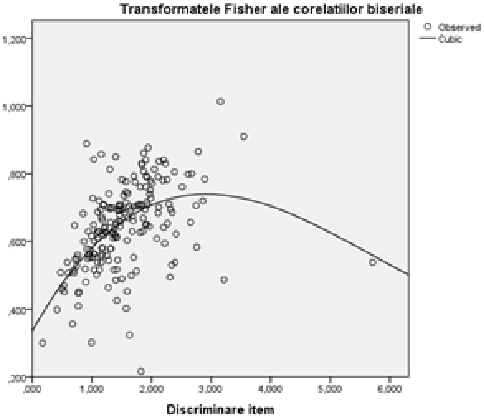


Figura V-36 Relația cubică dintre discriminarea itemilor clasici și cea a itemilor IRT în cazul întregului instrument

Tabel V-211 Estimarea relațiilor dintre cele două modele de discriminare în cazul întregului instrument

Model	Sumar model					Estimare parametri model			
	R Pătrat	F	df1	df2	Sig.	Constanta	b1	b2	b3
Linear	,181	41,022	1	186	,000	,516	,082		
Logaritmic	,267	67,782	1	186	,000	,590	,150		
Invers	,228	55,041	1	186	,000	,737	-,120		
Cvadratic	,290	37,836	2	185	,000	,379	,238	-,037	
Cubic	**,293**	**25,454**	**3**	**184**	**,000**	**,335**	**,314**	**-,073**	**,004**
Combinat	,156	34,331	1	186	,000	,513	1,140		
Putere	,250	61,909	1	186	,000	,576	,250		
S	,245	60,228	1	186	,000	-,295	-,214		
Creștere	,156	34,331	1	186	,000	-,667	,131		
Exponențial	,156	34,331	1	186	,000	,513	,131		
Logistic	,156	34,331	1	186	,000	1,949	,877		

The third hypothesis considers the same model of analysis, only we do not refer to discrimination but to the latent trait coverage. For the classical tests, the coverage level is given by the ratio of active response. This proportion, however, cannot be directly compared with the corresponding parameter of IRT items, requiring z score for normal distribution of proportion to one tail. It follows a probit indicator, comparable with the logit scale of IRT items parameter. To comply with strict compatibility between scales, the coverage level of IRT items has been transformed, also, in probits. The amplitude of distribution for IRT items is much higher in comparison with classic items, the average hovering around the middle of the continuum of latent trait, slightly to higher values, generally no significant differences between means. Standard errors of estimate are small; standard deviations were, again, higher for IRT items. This shows that the items assessing overall in the same area, the results can be compared. The fact is supported by the existence of significant and strong linear correlations, without significant differences. The cubic model is required again, the relationship between the two variables having the characteristics of an equation of the third degree.

Despite the difficulties, we have supported with real data what some researchers have shown through simulation studies. Item responses models are superior, estimators are more precise, much of the classical theory limits being exceeded. The answer to the original question is positive. Yes, we can estimate the amount of latent factor of a subject. This level of accuracy comes with a price however. The rigors are higher, mathematical mechanism is complicated, paper and pencil assessments cannot be done and the bank item needs to be extremely well designed.

Tabel V-222 Medii și abateri standard în cazul întregului set de itemi

		Probit (b) N=4043	**Z proporții răspuns activ N=323**
N	Valid	188	188
	Absent	0	0
Media		,18884	-,00056
Eroare standard medie		,088073	,031588
Abaterea standard		1,207603	,433110
Minimum		-2,297	-1,323
Maximum		3,798	1,041

Tabel V-223 Estimarea relațiilor dintre cele două modele ale nivelului factorului latent în cazul întregului set de itemi

Model	**Model**					**Sumar model**			
	R Pătrat	**F**	**df1**	**df2**	**Sig.**	**Constanta**	**b1**	**b2**	**b3**
Linear	,675	385,996	1	186	,000	,188	-2,290		
Invers	,022	4,100	1	186	,044	,202	-,005		
Cvadratic	,675	192,043	2	185	,000	,179	-2,285	,048	
Cubic	**,678**	**128,962**	**3**	**184**	**,000**	**,166**	**-2,473**	**,160**	**,385**

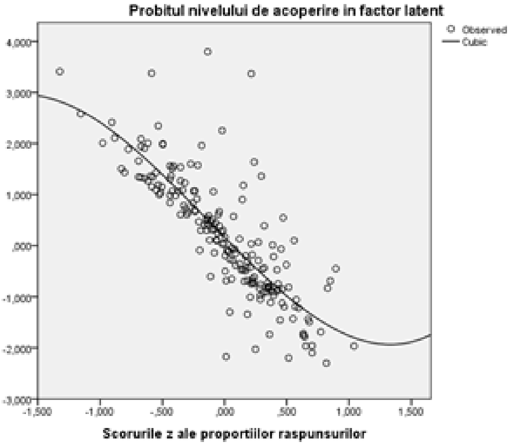


Figura V-42 Relația cubică dintre nivelul de acoperire în factor latent al itemilor clasici și cel al itemilor IRT în cazul întregului set de itemi

The chapter concludes with the limits of the research and the development perspectives. Since the second study tests contained a different number of items, we noted possible errors that can cause this difference, influencing the results. The origin of items from the classic tests leads to average levels of coverage in latent trait, which is another possible limitation. In the same category falls and the impossibility for most factors to comply to the measuring model assumption: there are differences between the characteristic curve of the model and the observed data.

Development perspectives consider several directions: from the study of multidimensional and polytomous models, to the design of a strong mechanism that identifies trends façade and controls random responses.

Chapter VI

Conclusions and discussion

The last chapter proposes a synthesis of the theory, practice and the research components of the thesis identifying the main elements presented in the paper.

Our intention was to provide a comprehensible summary encompassing the entire approach and mark the main results and concepts used.

References

1. Andersen, E. (1997). The rating scale model. În W. Van der Linden, & R. Hambleton, *Handbook of modern item response theory* (pg. 67-84). New York: Springer.

2. Baker, F. B. (1992). *Item response theory: Parameter estimation techniques.* New York: Marcel Dekker.

3. Baker, F. B. (2001). *The basics of item response theory.* Wisconsin: ERIC Clearinghouse on Assessment and Evaluation.

4. Bock, R. (1972). Estimating item parameters and latent ablility when responses are scored in two or more nominal categories. *Psychometrika*(37), 29-51.

5. Bock, R. D., & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443-459.

6. Bock, R., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*(35), 179-197.

7. Bock, R., & Mislevy, R. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*, 431-444.

8. Boekkooi-Timminga, E. (1991). A methos for designing Rasch model-based item banks. *Paper presented at the annual meeting of the Psychometric Society.* Princeton, NJ.

9. Constantin, T., & Macarie, A. (2012). *Chestionarul BigFive Plus - Manualul probei.* Draft, Universitatea Al. I Cuza, Facultatea de Psihologie și Științe ale Educației, Iași.

10. Constantin, T., Macarie, A., Gheorghiu, A., Iliescu, M., Fodorea, A., & Caldare, L. (2008). Chestionarul Big Five PLUS – rezultate preliminare. În M. Milcu, *Cercetarea Psihologică Modernă: Direcţii şi perspective* (pg. 46-58). București: Editura Universitară.

11. Costa, R., & McCrae, P. (2003). *Personality in Adulthood: A Five Factor Theory Perspenctive.* New York: Guilford Press.

12. Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Holt, Rinehart & Winston.

13. Davey, T., & Parshall, C. (1995). New algorithms for item selection and exposure control with computerized adaptive testing. *Anual meeting of the American Educational Research Association.* San Francisco, CA.

14. DeMars, C. (2010). *Item Response Theory.* New York: Oxford University Press.

15. Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists.* New Jersey, USA: Lawrence Erlbaum Associates, Publishers.

16. Finch, H., & Habing, B. (2007). Performance of DIMTEST- and NOHARM-Based Statistics for Testing Unidimensionality. *Applied Psychological Measurement*(31), 292-307.

17. Flaugher, R. (1990). Item Pools. În H. Wainer, *Computerized adaptive testing: A primer* (pg. 41-64). Hillsdale, NJ: Lawrence Erlbaum Associates.

18. Glas, G. A. (2002). Item calibration and parameter drift. În W. J. van der Linden, & G. A. Glas, *Computerized Adaptive Testing: Theory and Practice* (pg. 183-199). New York: Kluwer Academic Publichers.

19. Goldberg, L. (1999). A Broad-Bandwidth, Public-Domain, Personality Inventory Measuring the Lower-Level Facets of Several Five-Factor Models. *Personality Psychology in Europe, 7*, 7-28.

20. 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), 38-47.

21. Hambleton, R., Swaminathan, H., & Rogers, J. H. (1991). *Fundamentals of Item Response Theory.* London: Sage Publications Inc.

22. Keller, L. A. (2000). *Ability estimation procedures in Computerized Adaptive Testing.* American Institute of Certified Public Accountants.

23. Kingsbury, G., & Zara, A. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in EWducation*(2), 359-375.

24. Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

25. McDonald, R. (1967). Nonlinear factor analysis. (W. B. Press, Ed.) *Psychometric Monographs*(15).

26. Ostini, R., & Nering, M. L. (2006). *Polytomous Item Response Theory Models.* Thousand Oaks, California: Sage Publications.

27. Reckase, M. D. (2009). *Multidimensional Item Response Theory.* New York: Springer.

28. Reckase, M. D. (2010). Designing item pools to optimize the functioning of a computerized adaptive test. *Psychological Test and Assessment Modeling, 52*(2), 127-141.

29. Samejima, F. (1996). The graded response model. În W. Van der Linden, & R. Hambleton, *Handbook of modern item response theory.* New York: Springer.

30. Stan, A. (2002). *Testul psihologic. Evoluție, construcție, aplicații.* Iași: Polirom.

31. Stocking, M., & Swanson, L. (1998). Optimal design of item banks for computerized adaptive tests. *Applied Psychological Measuremen*(22), 271-280.

32. van de Linden, W. J., & Pashley, P. J. (2002). Item selection and ability estimation in Adaptive Testing. În W. J. van der Linden, & G. A. Glas, *Computerized Adaptive Testing* (pg. 1-25). New York: Kluwer Academic Publishers.

33. Van Der Linder, W. J., & Hambleton, R. K. (1997). *Handbook of Modern Item Response Theory.* New York: Springer.

34. Wesman, A. (1971). Writing the test item. În R. Thorndike, *Educational measurement.* Washington D.C.: American Council on Education.

35. Wilson, M. (2005). *Constructing Measures: An Item Response Modeling Approach.* New Jersey: Lawrence Erlbaum Assoicates Publishers.

36. Xitao, F. (1998, June). Item response theory and classical test theory: an empirical comparison of their item/person statistics. *Educational and Psychological Measurements, 58*(3), 357-373.