SBORNÍK PRACÍ FILOZOFICKÉ FAKULTY BRNĚNSKÉ UNIVERZITY
STUDIA MINORA FACULTATIS PHILOSOPHICAE UNIVERSITATIS BRUNENSIS
P 14 / 2010

MARTIN JELÍNEK, PETR KVĚTON, DALIBOR VOBOŘIL

ASSESSMENT OF RESPONSE PATTERN ABERRANCY IN EYSENCK PERSONALITY INVENTORY

Abstract
The presented study deals with two well standardized and relatively powerful indices based on Item Response Theory – z_3 and F_2 – which we used for detecting aberrant responding in the field of personality assessment. The indices were used on the Neuroticism and Extraversion scales from Eysenck Personality Inventory. The research sample consisted of 427 subjects. Both indices were computed for the whole sample and it was found that they yielded similar results. We selected subjects with the lowest z_3 index (highest aberrancy) values and further analyzed their response patterns. Our results suggest that in some cases the inconsistency was caused by faulty or careless responding, but in other cases by discrepancy between the subject's test behavior (item responding) and the theoretical construct being measured.

Keywords
measurement, personality, psychometrics

Assessment of Response Pattern Aberrancy in Eysenck Personality Inventory

Most psychological assessments stand on interpretations of questionnaires. But there is no secret that validity of questionnaires' inferences is often disputed. In one case a tested person can respond carelessly or randomly at all, in another case the tested person gives answers truthfully but his/her answers are not consistent with a normative pattern according to the theoretical background of the method. According to Reise and Waller (1993), at least three factors influence responding consistency: a) imperfection shared by all probabilistic measurement systems; b) faulty item responding (careless, misreading, etc.); c) individual differences to the normative theoretical expectations. Factor c) can be considered as lack of traitedness which refers to the agreement between subject's behavior and the construct measured.

Inconsistencies in respondent's answers can be psychometrically detected using various indices. Basically these indices are based either on Classical test the-

ory (CTT) or on Item response theory (IRT). The general approach to the problem in CTT can be demonstrated on Guttman's ideas. Consider a test consisting of several items. When answered by a relatively large sample we can sort the items by their difficulties (based on proportion correct score – p) in ascending order. A perfectly consistent individual pattern will then look like a line of 1s (correct answers) followed by 0s (wrong answers) and is called Guttman vector. Guttman error is such a case when any 1 is right to any 0 in the sorted line (i.e. subject is able to correctly answer a certain item but fails on a less difficult one). Pattern consisting of maximum number of Guttman errors (all 0s right to the 1s) represents Guttman reversed vector. This idea gave a ground for many other indices like Sato's Caution Index C, or Tatsuoka and Tatsuoka's Norm Conformity Index NCI (Meijer, Sijtsma, 1994).

In order to examine the validity of an individual response pattern in the IRT framework, various person-fit (PF) indices are used. Although a number of indices using different computation methods have been developed (see e.g. Tatsuoka, 1996; Drasgow, Levine, McLaughlin, 1987, etc.), their principle will always rely on the degree of consistence of a given response pattern and a valid pattern based on the relevant IRT model (Embretson, Reise, 2000). Advantageously, the PF indices based on IRT do not require the tested persons to go through the whole test with exact number of items and thus may be useful e.g. in the field of computerized adaptive testing (CAT).

Although many indices have been proposed in the scientific texts, there are two of them that are well standardised (their values do not systematically vary across different levels of θ) and show sufficient power for detecting aberrant response patterns (Drasgow, Levine, McLaughlin, 1987). Better known is the index originally proposed by Drasgow, Levine, Williams (1985) which was originally known as the z_3 index but others refer to it also as the Z_L index (Embretson, Reise, 2000) or l_z (Meijer, Sijtsma, 1994). It has been discussed that this index is also computably extendable for use with polytomous items (Drasgow, Levine, Williams, 1985). The latter one is F_2 index (Rudner, 1983). These two indices will be further discussed.

The logic of calculation

The core of IRT is the model of item responding. It allows determining a way in which the tested person – assuming we know his/her level of ability – is likely to respond to an item. Intuitively it applies that e.g. a very able person should resolve a very easy item with almost 100% certainty.

The person-fit indices are obtained in the following way: when the tested person has completed the whole test and we have thereby gained the information about the person's level of ability, we shall review the individual items and determine at each response whether it corresponds with his/her ability. The level of correspondence of the ability and the response is expressed by means of likelihood. It is then possible to determine the general credibility for the whole set of responses (i.e. test).

Calculation of z_3

The calculation (Drasgow, Levine, Williams, 1985) is based on the likelihood of a specific observed response pattern of a concrete subject expressed in the logarithmic form, i.e. as follows:

$$l_o = \sum_{i=1}^{n} u_i \log P_i(\hat{\theta}) + (1 - u_i) \log Q_i(\hat{\theta})$$

where i denotes the concrete item from the sample of n items, *log* is the natural logarithm, $\hat{\theta}$ = estimated ability, u_i = concrete response by a proband (1 – correct response, 0 – wrong response), P_i = probability of the correct response and $Q_i = 1\text{-}P_i$ (probability of the wrong answer). Probability of the correct response is computed as follows:

$$P_i(\hat{\theta}) = c_i + (1 - c_i) \frac{1}{1 + e^{-Da_i(\hat{\theta} - b_i)}}$$

where a_i denotes item discrimination parameter, b_i difficulty, and c_i guessing. When c_i=0 this three-parameter model is equal to two-parameter model with guessing parameter omitted. D is a constant equal to 1.7, which makes the logistic function close to the normal ogive function.

The expected value of l_o can be computed as:

$$E_3(\hat{\theta}) = \sum_{i=1}^{n} P_i(\hat{\theta}) \log P_i(\hat{\theta}) + Q_i(\hat{\theta}) \log Q_i(\hat{\theta})$$

In comparison to l_o, concrete response u_i is replaced by the probability of a correct response.

The standardized z_3 index is then defined as:

$$z_3 = \frac{l_o - E_3(\hat{\theta})}{\sigma_3(\hat{\theta})}$$

where conditional standard deviation:

$$\sigma_3(\hat{\theta}) = \sqrt{\sum_{i=1}^{n} P_i(\hat{\theta}) Q_i(\hat{\theta}) \left[\log\left(P_i(\hat{\theta}) / Q_i(\hat{\theta}) \right) \right]^2}$$

If the total log-likelihood of a specific subject's response pattern (l_o) equals its expected value for a given level of the latent trait (E_3), then the z_3 index value is zero. When z_3 is positive, the credibility of responses is higher than that predicted by the model. The unlikely response patterns are indicated by high negative index values. Considering that the index distribution should be roughly consistent with the standard normal distribution with an average of 0 and a standard deviation of

1, we can set e.g. a -2.33 value as an approximate value separating 1% of the most unsuitable patterns in the population.

The mathematic sequence of calculation of F_2
This index assesses person-fit by determining the degree of deviation from the expected response based on the latent trait level, summarized across items. The formula (Rudner, 1983) is as follows:

$$F_2 = \frac{\sum_{i=1}^{n}\left[u_i - P_i(\hat{\theta})\right]^2}{\sum_{i=1}^{n} P_i(\hat{\theta})Q_i(\hat{\theta})}$$

F_2 can be considered as discrepancy scores and so larger values indicate inappropriate response patterns. Given that, we can assume negative and very high correlation with z_3 index.
In the presented study we apply the procedures to data from Eysenck Personality Inventory with the aim to 1) describe and compare z_3 and F_2 indices, 2) identify persons with low credibility of response pattern and suggest possible reasons for that.

Method

Instrument

The Eysenck Personality Inventory, Czech version (Miglierini, Vonkomer, 1979): the questionnaire identifies two basic personality traits, extraversion and neuroticism. Each scale contains 24 dichotomous items; the respondent expresses his/her agreement or disagreement with a given statement by Yes/No response. Both scales provide sufficient levels of reliability ($KR20_E$=0.78; $KR20_N$=0.81). The inventory includes also a lie scale consisting of 18 items. Items from all scales are mixed together and some of them are reversed (with the exception of neuroticism scale).

Sample

The data comes from the research by Blatný, Osecká and Hrdlička (1998). The sample comprised 427 grammar school students from various towns of the South-Moravian Region (57% of women; average age of 16).

Data analysis

We calibrated the items of the neuroticism and extraversion scales with two parameter logistic model (2PL) in the Bilog 3.11 (Mislevy, Bock, 1997) software using marginal maximum likelihood estimation method. Trait level scores (θ) for all subjects were estimated using maximum likelihood method. Together eight

subjects (four in Neuroticism scale and four in Extraversion scale) with perfect negative or positive profiles were omitted from analysis because finite estimate of θ could not be reached. The mean and variance of θ were fixed at 0.0 and 1.0, respectively. Using items characteristics (b and a), individual trait levels and item responses, we were able to compute z_3 and F_2 indices for all subjects in our sample. For those interested, we give our interactive spreadsheet at http://www.psu.cas.cz/cato/z3.html free for download.

Results

The items of extraversion and neuroticism scales were calibrated using 2PL model because we found out that 1PL model would be too restrictive (item discrimination parameters, as can be seen in table 1, are diversified) and 3PL model is suitable only if one could expect guessing aspect in answering behavior. We present obtained item parameters in table 1.

TABLE 1. Item Parameters for EPI Scales

neuroticism				extraversion			
# in EPI		b	a	# in EPI		b	a
2	n1	-0.99	0.65	1	e1	-0.93	0.78
5	n2	0.10	0.48	3	e2	1.68	0.39
8	n3	-0.59	0.59	6	e3	1.75	0.45
10	n4	-1.10	0.52	9	e4	-0.32	0.49
13	n5	-1.65	0.43	11	e5	1.07	0.30
16	n6	-1.08	0.80	15	e6	-0.79	0.42
19	n7	1.60	0.62	18	e7	-0.80	1.16
22	n8	-1.29	0.71	20	e8	-0.72	1.31
25	n9	-1.76	0.55	24	e9	2.30	0.80
27	n10	0.20	0.86	26	e10	1.27	0.22
30	n11	0.40	0.93	29	e11	-1.74	0.92
32	n12	-1.74	0.43	31	e12	-0.40	0.92
37	n13	-0.81	0.50	35	e13	-0.50	1.24
39	n14	-0.49	0.70	38	e14	-0.93	0.51
41	n15	1.20	0.72	40	e15	0.13	0.15
45	n16	0.73	0.42	44	e16	0.42	0.34
47	n17	-0.01	0.8	46	e17	0.73	0.51
50	n18	2.01	0.63	48	e18	-2.43	0.21
53	n19	1.06	0.63	51	e19	1.55	0.60
55	n20	0.27	0.79	54	e20	0.03	0.70
58	n21	0.39	0.40	57	e21	-0.03	0.45
60	n22	0.74	0.92	59	e22	-1.21	0.79
64	n23	0.16	0.27	62	e23	0.17	0.90
66	n24	2.08	0.61	65	e24	-0.54	0.31

Description and comparison of z_3 and F_2 indices

Table 2 presents descriptive statistics for both scales (extraversion and neuroticism) and both indices. z_3 shows mean value around zero and standard deviation around 1 (z-distribution), while F_2 shows mean value around 1 and standard deviation 0.2.

TABLE 2. Descriptive Statistics for F_2 and z_3 Indices

	z_{3n}	z_{3e}	F_{2n}	F_{2e}
Mean	0.08	0.09	0.98	0.99
Median	0.21	0.19	0.96	0.96
Std. deviation	0.94	0.97	0.20	0.20

Note. z_{3n} and z_{3e} – z_3 indices for neuroticism and extraversion; F_{2n} and F_{2e} – F_2 indices for neuroticism and extraversion

Further analysis revealed that both indices are almost equivalent (r_e=-.96; r_n=-.98). Therefore, further results will present only one of the indices (z_3).

FIGURE 1. Distribution of z_3 Index for Neuroticism and Extraversion

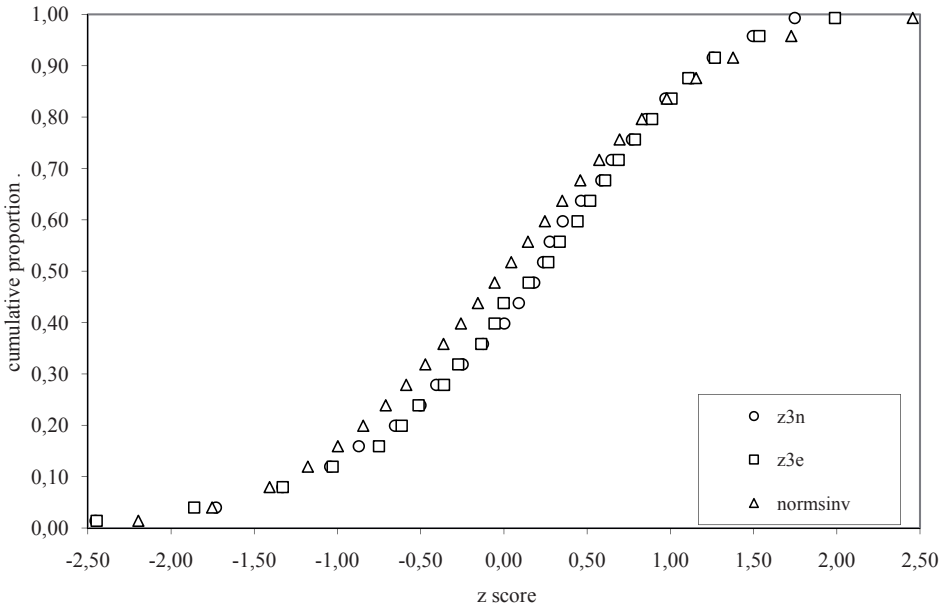

Figure 1 displays a slightly negatively skewed distribution of z_3 indices for both EPI scales. Lower number of values is located below the zero value than above zero, which is consistent with comparison of mean and median in table 2. It is also important to note that z_3 index does not correlate (r_n=-.02; r_e=.05) with θ and

thus measure of inconsistency level is not confounded by the overall θ value. Moreover, we found no relation between z_3 indices for both scales (r=.09).

Identifying persons with low credibility of response pattern

Table 3 and 4 bring a lot of useful information. The z_{3n} and z_{3e} lines show the essential information about the credibility of individual patterns (accompanied by values of the other scale index). Grayed fields in the tables clearly differentiate those five cases with high negative values of the indices from the ones with high positive values.

TABLE 3. Five Lowest and one Highest z_{3n} Scores with Associated Response Patterns

	Subject											
	# 161		# 124		# 419		# 413		# 338		# 118	
Item	Resp.	P	Resp.	P	Resp.	P	Resp.	P	Resp.	P	Resp.	P
n9	0	.91	0	.92	1	.79	1	.88	0	.71	1	.81
n12	0	.85	1	.87	0	.73	0	.82	1	.67	1	.76
n5	1	.85	1	.87	0	.72	1	.81	0	.65	1	.75
n8	0	.91	0	.93	0	.76	0	.88	1	.64	1	.79
n4	1	.83	0	.86	0	.66	1	.78	0	.57	1	.69
n6	1	.92	1	.94	1	.73	0	.87	1	.60	1	.78
n1	1	.86	1	.89	0	.67	1	.81	0	.55	1	.71
n13	0	.78	1	.81	1	.60	1	.73	0	.50	1	.63
n3	1	.78	1	.82	0	.56	0	.72	1	.45	1	.60
n14	0	.80	0	.84	0	.54	1	.72	0	.41	1	.59
n17	1	.72	0	.78	1	.39	0	.61	0	.26	1	.45
n2	1	.62	1	.66	1	.41	0	.55	0	.32	0	.44
n23	1	.56	0	.58	0	.44	1	.52	0	.39	0	.46
n10	0	.67	1	.74	1	.31	0	.55	0	.19	0	.37
n20	1	.63	1	.70	1	.30	1	.52	0	.19	0	.35
n21	1	.55	1	.59	0	.38	1	.49	1	.31	0	.41
n11	1	.61	1	.69	0	.23	1	.47	0	.13	0	.29
n16	0	.49	1	.53	0	.32	1	.43	0	.25	0	.34
n22	1	.48	1	.57	0	.15	1	.35	0	.08	0	.19
n19	1	.40	0	.46	0	.18	0	.31	1	.12	0	.21
n15	0	.35	1	.41	1	.13	0	.26	0	.08	0	.16
n7	1	.28	1	.33	0	.11	1	.21	1	.07	0	.13
n18	0	.19	0	.24	0	.07	0	.14	0	.05	0	.09
n24	1	.19	1	.23	1	.07	1	.14	1	.05	0	.09
θ_n	0.68		0.91		-0.35		0.33		-0.80		-0.17	
z_{3n}	-3.39 (z_{3e}=-0.58)		-3.11 (z_{3e}=-0.57)		-3.05 (z_{3e}=0.27)		-3.05 (z_{3e}=-0.64)		-2.65 (z_{3e}=0.08)		2.34 (z_{3e}=-0.13)	

Note. Grayed fields designate unexpected (by means of item response functions) answers. Items are ordered by ascending difficulty. Resp. – subject's responses (1 keyed; 0 nonkeyed); P – expected response probability.

TABLE 4. Five Lowest and one Highest z_{3e} Scores with Associated Response Patterns

	Subject											
	# 95		# 251		# 2		# 353		# 96		# 147	
Item	Resp.	P	Resp.	P	Resp.	P	Resp.	P	Resp.	P	Resp.	P
e18	0	.68	1	.74	0	.69	1	.63	0	.61	1	.66
e11	0	.90	1	.97	1	.92	1	.80	0	.71	1	.86
e22	0	.77	0	.91	1	.8	0	.62	0	.51	1	.70
e14	1	.63	1	.78	0	.66	0	.52	0	.45	1	.57
e1	1	.69	1	.87	1	.73	0	.53	1	.42	1	.61
e7	0	.72	1	.93	0	.77	0	.48	0	.32	1	.60
e6	0	.58	1	.72	1	.61	1	.49	0	.43	1	.54
e8	1	.71	0	.94	1	.77	0	.43	0	.27	1	.58
e24	0	.53	1	.64	1	.55	0	.46	0	.42	1	.49
e13	0	.60	0	.90	1	.67	0	.32	0	.19	0	.45
e12	0	.53	1	.81	0	.59	1	.33	1	.23	0	.43
e4	1	.50	0	.67	1	.53	1	.39	0	.33	0	.44
e21	1	.45	1	.61	0	.47	0	.35	1	.29	0	.39
e20	1	.40	1	.64	0	.44	0	.26	0	.19	0	.32
e15	1	.47	0	.53	0	.48	0	.44	0	.42	1	.45
e23	1	.32	1	.63	0	.37	1	.17	1	.11	0	.24
e16	1	.40	0	.52	0	.42	0	.33	0	.29	0	.36
e17	1	.29	1	.46	0	.32	1	.20	1	.16	0	.24
e5	1	.33	1	.43	1	.34	1	.27	1	.24	0	.30
e10	1	.36	0	.43	0	.37	0	.31	0	.28	0	.33
e19	1	.13	1	.26	0	.15	0	.08	0	.06	0	.10
e2	1	.21	1	.32	1	.23	1	.16	1	.13	0	.18
e3	0	.17	1	.28	1	.19	1	.12	0	.10	0	.14
e9	0	.03	1	.08	1	.03	0	.01	0	.01	0	.02
θ_e	-0.32		0.53		-0.17		-0.84		-1.18		-0.59	
z_{3e}	-4.87		-4.15		-2.69		-2.67		-2.62		2.25	
	(z3n=-2.18)		(z3n=0.30)		(z3n=-0.13)		(z3n=N/A)		(z3n=-1.93)		(z3n=-0.98)	

Note. Grayed fields designate unexpected (by means of item response functions) answers. Items are ordered by ascending difficulty. z_{3n} score for subject ID 353 is assigned N/A (= not available) because of a positive infinite estimation of theta. Resp. – subject's responses (1 keyed; 0 nonkeyed); P – expected response probability.

Graphical depiction of unexpected responses helps us to identify most problematic items where all our "inconsistent" subjects give responses totally against the model probability. We identified one such item in case of neuroticism scale (n24: "Do you suffer from sleeplessness?") and two in case of extraversion scale (e2: "Are you usually carefree?" and e5: "Would you do almost anything for a dare?").

Discussion

It is evident that psychological testing, and particularly the testing of mass character, is affected, apart from other sources of errors, also by the kind of errors stemming from the very behavior of persons within the testing procedure. PF indices can be used in the area of performance testing to identify persons cheating in the test, but also to identify persons with specific abilities or diagnose cognitive errors of the concrete subjects (Tatsuoka, Tatsuoka, 1983).

Interesting is also the use of PF indices in personality testing as presented in our study. The personality research based on the questionnaire methods is faced with the problem of identification of faulty answers, particularly in certain groups of subjects such as adolescents (Handel et al. 2006). Another source of bias can be lack of, so called, traitedness. In general, measurement methods (like questionnaire EPI) stand on nomothetic trait construct and are usually constructed using advanced statistical procedures (factor analysis). It is obvious that this approach does not allow us to expect that every single individual will perfectly fit to the construct.

The way of interpreting PF indices depends on the aspirations. Researcher working with large amounts of data can use them to clear his/her dataset from uninterpretable (inconsistent with the model) records without looking after the causes. Another situation arises in clinical setting. A skilled psychological professional knows that the same questionnaire score does not always mean the same quality for two different clients. Using a PF index he/she can assure that with sufficient value of the index, the score is interpretable by means of the theory behind the method. But when non-interpretable score is found, he/she should ask for the reasons. From this point of view, there may be two main types of inconsistencies: a) client did not understand item contents, misread items, or e.g. filled out the answer sheet carelessly (faulty answering); b) client honestly filled out the questionnaire but the theoretical construct for some reason does not apply to him/her and there is a need for additional information to explore.

In our study we identified several subjects with highly inconsistent answering in scales of extraversion and neuroticism. Our goal was not only to identify them but also try to suggest possible reasons for it. Reise and Waller (1993) explored data obtained using Multidimensional Personality Questionnaire (MPQ). For distinction between faulty responding and lack of traitedness they used 1) several validity scales incorporated in MPQ (e.g. Variable Response Inconsistency Scale, which is based on comparison of answers on items with similar content) and 2) comparison of PF index values between different scales. The EPI questionnaire used in our study contains Lie Scale which is focused rather on social desirability than on consistency of responding and thus is not suitable for our purposes. Therefore, we had only one clue for distinction between faultiness and lack of traitedness – comparison of z_3 index scores from the two scales. When the index indicates inconsistency in both scales, then we assume that it is faultiness that plays a major role here. Such cases are ID # 95 and # 96 as shown in table 4.

When the inconsistent tendency was found only in one of the scales, we inclined to interpret it rather as a lack of traitedness.

It was also interesting to look on items with a high rate of highly unexpected responses. Such an item can be found in the Neuroticism scale (n24: "Do you suffer from sleeplessness?"). Although sleeplessness indeed is an important indicator of neuroticism, there surely exist people that are sleepless from other reasons than that (e.g. neurological disorders).

Even though the PF indices were found useful in many areas (discussed above), also less promising attempts for application can be found in the relevant literature. For example, Brown and Harvey (2003) tried to identify faking in Five Factor Model personality test (Conscientiousness and Agreeableness scales) but with no substantial success. As it is well known from practice, motivated individuals are often capable of pretending desired (in his/her opinion) characteristics and doing it consistently.

References

Blatný, M., Osecká, L., & Hrdlička, M. (1998). Zdroje sebehodnocení u temperamentových typů [Sources of self-esteem in temperament types]. *Československá psychologie 42,* 297-305.

Brown, R. D., & Harvey, R. J. (2003). Detecting personality test faking with appropriateness measurement: fact or fantasy? Paper presented at the 2003 Annual Conference of the Society for Industrial and Organizational Psychology, Orlando.

Drasgow, F., Levine, M. V., & McLaughlin, M. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement 11,* 59-79.

Drasgow, F., Levine, M. V., & Wiliams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology 38,* 67-86.

Embretson, S. E., Reise, S. P. (2000). *Item response theory for psychologists.* London: Lawrence Erlbaum Associates.

Handel, R. W., Arnau, R. C**.,** Archer, R. P., & Dandy, K. L. (2006). Evaluation of the MMPI-2 and MMPI-A True Response Inconsistency (TRIN) scales. *Assessment 13*, 98-106.

Meijer, R. R., & Sijtsma, K. (1994). Detection of Aberrant Item Score Patterns: A Review of Recent Developments. *Research Report, Faculty of Educational Science and Technology, University of Twente 94*, 3-26.

Miglierini, B., & Vonkomer, J. (1979). *Eysenckov osobnostný dotazník – EOD* [Eysenck Personality Inventory – EPI]. Bratislava: Psychodiagnostické a didaktické testy.

Mislevy, R. J., & Bock, R. D. (1997). *Bilog 3.11.* Mooresville, IN: Scientific Software International.

Reise, S. P. & Waller, N. G. (1993). Traitedness and the assessment of response pattern scalability. *Journal of Personality and Social Psychology 65,* 143-151.

Rudner, L. M. (1983). Individual Assessment Accuracy. *Journal of Educational Measurement 20,* 207-219.

Tatsuoka, K. K. (1996). Use of generalized person-fit indices, zetas for statistical pattern classification. *Applied Measurement in Education 9,* 65-75.

The study was performed with the support of The Czech Science Foundation (project nr. 406/09/P284) and is part of Research project of the Institute of Psychology, Academy of Science of the Czech Republic, identification code: AV0Z70250504.