Psychology Science, Volume 46, 2004 (4), p. 466 - 476

On changing the position of items in personality questionnaires
Analysing effects of item sequence using IRT

TUULIA M. ORTNER[1]

Abstract

Although personality questionnaires are widely criticized, they still represent an essential instrument for psychological assessment. By referring to the methodological malpractice of the extraction of scales and changes in item position, this paper deals with test theoretical changes in the quality of questionnaires that changes to the item position have. In the following study, two experimental groups filled out two versions of the EPP-D (Eysenck, Wilson, & Jackson, 1998): One consisting of Rasch-homogenous items in its conventional order and one group in the exact reversed order. While statistical comparisons of mean scores attained in the two versions showed no significant differences between the two versions, analyses with IRT showed different difficulties for items in three of seven scales which had been caused by the item order. It is inferred that classical test theoretical approaches lack information and are not sufficient for the study of effects of changed item orders.

Key words: item response theory, item order, questionnaire, adaptive testing

[1] Tuulia M. Ortner, Division of Assessment and Applied Psychometrics, Faculty of Psychology, University of Vienna, Liegiggasse 5, A-1010 Vienna, Austria; E-mail: tuulia.ortner@univie.ac.at

1. Introduction

At present, personality questionnaires still represent an essential instrument for psychological assessment. Nevertheless, problems of their use such as faking, psychometrical problems or even problems of reasonableness are well known (Kubinger, 2002).

In the construction of questionnaires, items are normally presented from a mix of the scales in a random order. The aim of this practice is to reduce obviousness and fakeability. Changes to the contents of a questionnaire are supposed to keep the motivation of the tested person high (Boecker, Keil, Eiser, & Kline, 1987) and the effects of context low. Anger (1969) even mentions effects of *extinction*: A new question or topic is supposed to tend to minimize the effects of priming.

However, in the use of questionnaires, one problematic methodological malpractice is well known: The extraction of single scales out of multidimensional personality questionnaires, which is mostly unavoidably combined with changes to the original item order. One reason for scale extraction is given by setting the focus on just one single or a few special traits from a questionnaire. Other causes for single item extraction might be given by the computerized adaptive presentation of a questionnaire: With adaptive computer tests, the examinee's ability level can be iteratively estimated during the testing process and items can be selected based on the current ability estimate. Each person is then only confronted with the items of most interest (Van der Linden & Glas, 2000; Wainer et al., 1990; Weiss, 1982). Changes to the item order are, in this case, unavoidable.

Nevertheless, several studies have already focussed on the problem of changes to the item order. Most of them deal with problems that occur by blocking item scales: Rost and Hoberg (1997) mixed items according to the self concept of school performance into distractor items a) by randomly or b) in evidently distinguishable blocks. Analysis according to classical test theory showed no differences in the construct validity regarding factor structure between the two forms. However, in the majority of the cases, there was an increase in homogeneity and the average scores were higher in the cases of item blocking. No differences in the randomized and grouped item presentation with regard to reliability and validity were shown for scales measuring occupational and life satisfaction (Schriesheim, Kopelman, & Solomon, 1989). Significant results on the level of average scores and internal consistence in a blocked versus conventional version were also shown for a multidimensional personality questionnaire (Krampen, Hense, & Schneider, 1992) and for an instrument measuring perceived competence and control beliefs (Krampen, 1993). However, the effects were mostly inconsistent due to the scales and were also mainly quite small. The problem of the presentation of items in a block-wise manner due to the effects of changed item orders is a special one and should be mentioned only marginally here: Additional effects due to the increased obviousness of the matter being addressed and therefore an increase in fakeability are also very likely in this special case.

In contrast, the effects of random changes to the item order of multidimensional questionnaires have not been part of many studies yet. Such studies have only been conducted a few times, e.g. Abel (2003) conducted a comparison of two different versions of an interest scale. The versions differed in item order (conventional versus reversed): Raw scores, item difficulties and total score-item correlations showed differences between the two versions.

Some reasons for different answer behaviour concerning items in relation to their position in a questionnaire do seem possible: One reason might be the increased attention at the start

of filling out a questionnaire which allows one to focus more strongly on one's personal intentions. This might also include the phenomenon of "self alertness". Items presented at the beginning may perhaps be of greater difficulty than those presented at the end of a questionnaire. On the other hand the examinees are also more likely to be drowsy towards the end, which also might cause certain tendencies in their answer behaviour, rigid behaviour or behaviour by chance due to lack of compliance (Arnold, Eysenck, & Meili, 1980). Another complex of behaviours that influence assumptions includes possible priming or halo-effects which are caused by certain items being presented before others, as was the case in the experimental induction by Krahé and Herrmann (2003). The effect of priming might also include another source of differences in answer behaviour which is a bit more complex: One idea is that a kind of image is also formed by the first few items presented in a questionnaire: e.g. very difficult items (in personality questionnaires: items with a low probability of agreement) at the beginning of a questionnaire might lead to the test person having the opinion that items are generally formulated rather extremely or they tend to be unreasonable and therefore the examinees change their behaviour.

2. Aim of the study

The studies mentioned above proved their hypotheses by applying conventional reliability and validity concepts but did not deal with item calibration analyses which IRT (Item Response Theory) approaches enable (Molenaar, 1995). However, a new approach seems to be essential for various reasons. Firstly, the application of IRT to personality questionnaires is necessary if one plans to present different items out of an item pool to different test persons. So, one of the most important reasons for item extractions and changes to the item order is the application of computerized adaptive testing. The conventional methods of classical test theory, which implies the summation of the score, must fail in this case because each person has answered different questions. The amount of "solved" items is obviously not a fair criterion in adaptive testing. Only in the case of the Rasch-model fit of a given item pool is the predicted trait levels equivalent estimable regardless of the items on which it is based. Secondly, because the order of items presented might be different for every tested person, local stochastic independence is required, this being one of the provable implications of the Rasch-model.

This paper presents an experiment: One sample of persons (control group) answers a questionnaire in the conventional order. Another sample, the experimental group, responds to the same questionnaire in a reversed item order, as was the case in former studies conducted.

The particular aim of the following study is the application of an IRT model for the analysis of equableness and item homogeneity of two different versions of the same questionnaire, differing only in item order. Similar analyses were carried out by Dissauer (1979) and Hahne (1999) for achievement tests. It was shown that particular items are more difficult at the beginning of a test series than at the end.

The question is now, whether the change to the item order of a personality questionnaire leads to different estimations of item parameters for the two different versions.

3. Questionnaire

A German pre-version of the translated *Eysenck Personality Profiler* (EPP; Eysenck, 1995), existing of all 440 items was used as the personality questionnaire in this study. It is published with reduced amount of items by Eysenck, Wilson and Jackson (1998). It has been widely used in English speaking countries in research and consultancy. The Eysenck Personality Profiler measures 21 traits of personality which are consistent with the three major dimensions of personality as defined by Hans-Jürgen Eysenck.

According to the dimension *Extraversion*, the traits are: *Activity, Sociability, Expressiveness, Assertiveness, Ambition, Dogmatism* and *Aggressiveness*. The *Neuroticism* Scales include *Inferiority, Unhappiness, Anxiety, Dependence, Hypochondria, Guilt* and *Obsessiveness*. The dimension of *Psychoticism* subsumes *Risk-taking, Impulsivity, Irresponsibility, Manipulativeness, Sensation-seeking, Tough-mindedness* and *Practicality*. A lie scale is also included. The test was adapted to the German language by Bulheller and Häcker (1998). The published questionnaire includes 176 items and is to be answered with “yes”, “I don’t know” or “no”.

Perfahl (1998) analysed the original German translated version of the EPP (consisting of 440 items) for a model fit with regard to IRT. She tested a sample of 349 persons (142 male, 207 female) aged 17 to 77 (median = 34) with regard to the Partial Credit Model (Masters, 1982). The score was used as the splitting criterion. Model fit was assumed if at least fourteen of twenty items per scale showed fit. Other scales were excluded as they did not fit. Model fit was shown for ten (out of 22) scales.

4. Dichotomization and Reanalysis of Perfahl’s Data (1998)

As mentioned above, the answer form of the EPP makes it possible for persons to answer the presented questions in three categories, with a middle category ”I don’t know”. Problems may occur here because an examinee might choose the middle option for various reasons: This may be that the person does not understand the particular question, does not know him- or herself this well, has no interest in answering the questions, or perhaps understands this as a middle parameter value (Moosbrugger, Fischbach, & Schermelleh-Engel, 1998). In view of this fact that the middle category leads to numerous problems (e.g. finding an appropriate probabilistic model) and according to future plans of constructing an adaptive version of the questionnaire, the calculation of answer categories was changed to a dichotomous format.

First, dichotomization of Perfahl’s data was carried out by calculating the answer category “I don’t know” as “no” or “no distinction” in all cases. If this did not lead to a Rasch-homogenous item pool, the misfitting items were dichotomized in a second step according to an expert rating accomplished by seven students (whose main course was psychological assessment). These students had evaluated the meaning of the category for each item. Only if six of the seven raters were of the same opinion regarding an item that option was taken into account. In all other cases it was counted as “no distinction”. In most cases, the decision of the raters indicated that “I don’t know” answers might be a way of answering in a more socially desirable manner than by using the extreme positions. However, in many cases these well-trained experts could also not agree on the reason lying behind for using that category.

The model used for the second analysis was the dichotomous Rasch Model (Rasch, 1960). Parameter estimates were performed by LPCMWin (Fischer & Ponocny-Seliger, 1998) using the Likelihood Ratio Test (LRT) by Andersen (1973). Only those scales which had already shown a model fit in Perfahl's analysis were included in the analyses. The sample was divided into two subgroups according to the raw score achieved by the examinees (high versus low). Only in cases of additional item removal were the criteria sex and age additionally used to confirm the result. This only happened in two cases.

As a first result, the expert rating could not turn those scales to model fit that had not already proved it during the "raw" dichotomization. The results of the model tests showed a fit for seven of the ten scales after dichotomization. These seven scales (*dogmatic*, *expressive*, *anxious*, *hypochondric*, *manipulative*, *depressed*, *impulsive*) are spread on the three dimensions as defined by Eysenck. As expected, the items of every scale were not distributed equally between the potential areas of difficulty; e.g. analyses showed that in some of the scales not many items for the mid range area of difficulty are available (e.g. in the scale *dogmatic*), and in other scales there are few or no items for the very difficult or very easy levels (e.g. in the scales *depressed*, *anxious*). So, after dichotomization, seven of ten scales were also identified as concurrent with the dichotomous model from Rasch. In total, between 14 and 16 model fitting items were left for further analysis per scale.

Results are shown in Table 1.

Table 1:
Results of LRTs, splitting criteria and the number of additionally removed and remaining items after the dichotomization of the data from Perfahl (1998)

Scale	Statistics	Splitting criteria	Additional Items removed	Items left
dogmatic	χ^2=13.45; df = 15	score	0	**16**
expressive	χ^2=26.57; df = 13	score	4	**14**
	χ^2=22.02; df = 15	age	4	
	χ^2=27.58; df = 13	sex	4	
assertive	χ^2=28.05; df = 13	score	-	**-**
anxious	χ^2=23.45; df = 14	score	0	**15**
hypochondrial	χ^2=22.77; df = 13	score	0	**14**
depressed	χ^2=16.49; df = 13	score	0	**14**
manipulative	χ^2=24.75; df = 13	score	0	**14**
risk taking	χ^2=25.32; df = 13	score	-	**-**
	χ^2=48.08; df = 14	age		
	χ^2=28.64; df = 13	sex		
impulsive	χ^2=23.15; df = 14	score	0	**15**
dissimulation	χ^2=40.53; df = 14	score	-	**-**

5. Experiment

The test versions were implemented in the program T·N·T (Brugger, unpublished). This program allows for the running of items and standardized scoring, as well as having a tool for adaptive testing and simulations. The weighted Maximum Likelihood Method from Warm (1989) was used for parameter estimations.

Two versions of the EPP-D were programmed:

- One "conventional" version (*control group*) existing of all the Rasch-homogenous items in the original order (that means all items translated from the original English version excluding the non Rasch-homogenous items)
- One "reversed" version existing of the same items as the version described above but in the reversed order (*experimental group*)

Because changes in model fit due to changes in answer format are possible, the same answer format as in original version was also used in this second study: Persons could answer in the three categories described above. Answers were scored dichotomously in the same way as the dichotomization was carried out.

6. Sample

The participants were young male Viennese who had been called up for military service. If they agreed (~80%), they were tested after the standardized psychological testing conducted by the Psychological Service of the Austrian Armed Forces. They were only tested if a) they were evaluated as being motivated by the conductor b) they had solved the standardized battery quickly without noticeable problems and no language problems were known. To reduce faking, the conductor pointed out that all results are handled anonymously and are not evaluated to determine the military appropriateness of the test persons.

A sample of 168 persons (all male) aged 18 to 34 (median = 18) was tested. The score was used as splitting criterion. Participants were well balanced concerning their educational status. 83 persons answered the questions in the standard order. 88 examinees gave their answers to the same items in the reversed order. For technical reasons, the persons were not assigned randomly to the two groups. However, the origin of the participants and the testing procedure were the same, and the conductors were not informed about differences between the questionnaires given out.

7. Results

For the seven scales, item parameters were estimated under the same conditions as described above.

By applying the dichotomous Rasch model with the common splitting criterion of "score", six of the seven scales showed a model fit: Only the scale "impulsive" failed. Results are shown in Table 2.

Table 2:
Results of LRTs using splitting criterion score (statistics printed in bold show significant results from the LRT)

Scale	Statistics	Splitting criterion
dogmatic	χ^2=30.57; df = 15	score
expressive	χ^2=22.54; df = 13	score
anxious	χ^2=11.41; df = 14	score
hypochondrial	χ^2=17.63; df = 13	score
depressed	χ^2=18.37; df = 13	score
manipulative	χ^2=16.35; df = 13	score
impulsive	**χ^2=51.15; df = 14**	score

In a second step, the criterion "test version" was used for splitting the sample size: There was no model fit for three of the seven scales with regard to the original item pool: However, for the scales *hypochondrial* and *depressed* a model fit was found after dropping one more item. Not even additional extraction of items leads to a model fit for the scale *dogmatic* (see Table 3).

Table 3:
Results of LRTs using splitting criterion test version (statistics printed in bold show a significant result)

Scale	Statistics	Splitting criterion
dogmatic	**χ^2=37.31; df = 15**	test version
expressive	χ^2=12.63; df = 13	test version
anxious	χ^2=12.48; df = 14	test version
hypochondrial	χ^2=14.42; df = 12*	test version
depressed	χ^2=21.99; df = 12*	test version
manipulative	χ^2=21.48; df = 13	test version
impulsive	χ^2=20.43; df = 14	test version

** model fit was achieved after removal of one additional item*

Because of the moderate number of persons tested, deficient assumptions of model fit were probable. For this reason, graphical model checks were also conducted for the criterion "test version".

The graphical model checks (Fig. 1) show that the estimated item parameters fit to the 45°degree line very well for the scales *anxious*, *expressive* and *impulsive*. All items almost lie on an imaginary line there. A fit from a graphical point is obviously not given, for the scales *dogmatic*, *hypochondrial* and *depressed*.

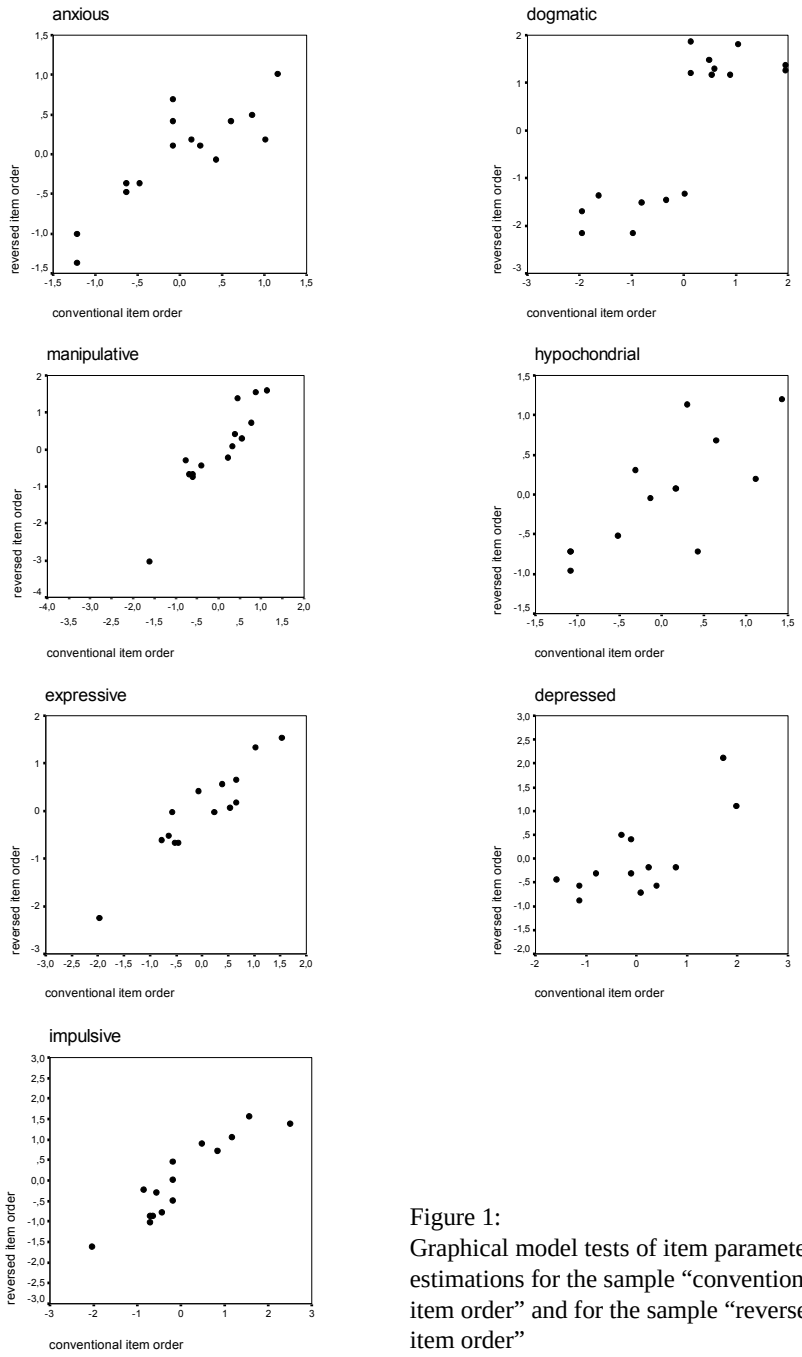

Figure 1:
Graphical model tests of item parameter estimations for the sample “conventional item order” and for the sample “reversed item order”

To lay attention additionally on one more regularly used attribute for evaluation of equivalence of different test versions, the scores achieved in the two studies were taken into account and statistically compared in the two experimental groups (t-Test): By adjusting alpha ($\alpha = .007$), no scale showed significant differences between the two test versions (see Table 4).

Table 4:
Results of *t*-Tests for independent samples for scores in the samples "conventional" item order and "reversed" item order

	version	**n**	**average score**	**variance**	***t***	**sig.**
dogmatic	conventional	83	6.16	2.02	-.759	.449
	reversed	88	6.41	2.31		
expressive	conventional	83	6.08	2.12	-1.683	.094
	reversed	88	6.65	2.25		
anxious	conventional	83	2.58	2.61	-1.084	.280
	reversed	88	3.02	2.74		
hypochondrial	conventional	83	1.42	1.59	-.821	.413
	reversed	88	1.63	1.65		
depressed	conventional	83	1.47	1.91	-2.454	.015
	reversed	88	2.30	2.44		
manipulative	conventional	83	4.94	2.36	1.158	.249
	reversed	88	4.52	2.35		
impulsive	conventional	83	6.17	2.70	.416	.678
	reversed	88	5.98	3.27		

8. Discussion

Two significant results were shown in this study. Firstly, various aspects of two methodological approaches regarding the analyses of the quality of personality questionnaires were compared: Analyses of the average scores were unable to uncover any metrical problems due to changes in item position in this case. Further results of the psychometric equivalence of the two reduced versions of the EPP-D were found by applying IRT: For seven of the model fitting scales, which had been recognised in an earlier study, an analysis with a new sample, using the splitting criterion *score*, showed one non fitting scale. Using the splitting criterion *test version*, three of the seven scales failed the model fit. This shows that the results cannot be interpreted as meaning that the item parameters are generally unstable in personality questionnaires: Using the conventional partition criterion, model fit concerning one sample for parameter estimation was mostly confirmed in a second sample.

The second result shows that a simple change to the item order leads to changes in item difficulties in three of seven scales. In this case, items in different positions of a test showed a change in their item difficulty.

For application in adaptive testing, these results show serious problems: A fair measurement of a person's tendency in answering different items does not seem reasonable at this point in time. However, these results refer at least to one questionnaire, the EPP-D, and just to the Rasch-model fitting items, not to the whole published version of this questionnaire. However, the psychometric quality of the excluded, not fitting items is at least uncertain: Such analyses of evaluation of psychometric quality and stability as done in this study are impossible for items not fulfilling minimal standards as claimed by IRT and therefore their quality remains doubtful. Nevertheless, the gained results refer to a modified version of the EPP-D, not the whole questionnaire and therefore a generalization of the results is restricted.

Results concerning other questionnaires should follow. With regard to the adaptive application of personality questionnaires, future results should also show which constraints lead to different estimations of item difficulties.

Acknowledgements

I would like to thank Erich Frise and Christof Brugger for support in the recruitment of test persons.

References

1. Abel, J. (2003). Testtheoretischer Vergleich von Versionen des Allgemeinen-Interessen-Strukturtests (AIST). [Test theoretical Comparison of Versions of the Allgemeinen-Interessen-Strukturtest (AIST)]. (Vol. 17). Saarbrücken: Sondersammelgebiet Psychologie an der Saarländischen Universitäts- und Landesbibliothek Saarbrücken.
2. Andersen, E. B. (1973). A Goodness of Fit Test for the Rasch Model. Psychometrika, 38, 123-140.
3. Anger, H. (1969). Befragung und Erhebung. [Questioning and Survey]. In C. F. Graumann (Ed.), Sozialpsychologie (Vol. 7/1). Göttingen: Hogrefe.
4. Arnold, W., Eysenck, H.-J., & Meili, R. (Eds.). (1980). Lexikon der Psychologie. [Lexicon of Psychology]. Freiburg: Herder.
5. Boecker, M., Keil, K. S., Eiser, R. J., & Kline, P. (1987). Are personality questionnaires answered mindlessly? European Journal of Personality, 1(4), 231-239.
6. Brugger, C. (unpublished). Teach and Test (TNT). Vienna.
7. Dissauer, G. (1979). Lern- bzw. Übungseffekte innerhalb von Testserien [Effects of Learning and Practice within Test Series]. Unpublished Dissertation, Vienna.
8. Eysenck, H.-J. (1995). The Eysenck Personality Profiler and Eysenck's Theory of Personality. London: Corporated Assessment Network Ltd.
9. Eysenck, H.-J., Wilson, C. D., & Jackson, C. J. (1998). Eysenck Personality Profiler (EPP-D). Franfurt/M.: Swets.
10. Fischer, G., & Ponocny-Seliger, E. (1998). LPCM-Win. Structural Rasch modeling. Handbook of the usage of LPCM - Win 1.0. Groningen: ProGamma.

11. Hahne, J. (1999). Lerneffekte innerhalb von Leistungstests. [Learning Effects whithin Achievement Tests.]. Unpublished Diploma Thesis, Vienna.
12. Krahé, B., & Hermann, J. (2003). Verfälschungstendenzen im NEO-FFI: Eine experimentelle Überprüfung. [Faking Tendencies in the NEO-FFI: An Experimental Test]. Zeitschrift für Differentielle und Diagnostische Psychologie, 24, 105-117.
13. Krampen, G. (1993). Effekte von Bewerbungsinstruktionen und Subskalenextraktion in der Fragebogendiagnostik. [Effects of Job Application Briefing and Extraction of Subscales in the Assessment with Questionnaires.]. Diagnostica, 39, 97-108.
14. Krampen, G., Hense, H., & Schneider, J. F. (1992). Reliabilität und Validität von Fragebogenskalen bei Standardreihenfolge versus inhaltshomogener Blockbildung ihrer Items. [Reliability and Validity of Questionnaire Scales in conventional Order versus content homogenous Blocking of Items.]. Zeitschrift für experimentelle und angewandte Psychologie, 39, 229-248.
15. Kubinger, K.-D. (2002). Psychology's challenge when personality questionnaires are applied for individual assessment. Psychologische Beiträge, 44, 3-9.
16. Molenaar, I. W. (1995). Some Background for Item Response Theory and the Rasch-Model. In G. H. Fischer & I. W. Molenaar (Eds.), Rasch Models. Foundations, Recent Developments, and Applications. New York: Springer.
17. Moosbrugger, H., Fischbach, A., & Schermelleh-Engel, K. (1998). Zur Konstruktvalidität des EPP-D. [About Construct Validity of the EPP-D]. In H.-J. Eysenck, C. D. Wilson & C. J. Jackson (Eds.), Eysenck Personality Profiler (EPP-D) (pp. 89-118). Frankfurt: Swets Test Service.
18. Perfahl, B. B. (1998). Eine testtheoretische Analyse des neuen EPP-D. [A test theoretical Analysis of the new EPP-D]. Unpublished Diploma Thesis, Vienna.
19. Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Nielsen & Lydiche.
20. Rost, D.-H., & Hoberg, K. (1997). Itempositionsveränderungen in Persönlichkeitsfragebogen: Methodischer Kunstfehler oder tolerierbare Praxis? [Change of Item Positions in Personality Questionnaires: Methodical Malpractice or tolerable Practice?]. Diagnostica, 43(2), 97-112.
21. Schriesheim, C. A., Kopelman, R. E., & Solomon, E. (1989). The effect of grouped versus randomized questionnaire format on scale reliability and validity: A three-study investigation. Educational and Psychological Measurement, 49, 487-508.
22. Van der Linden, W. J., & Glas, C. A. W. (Eds.). (2000). Computerized Adaptive Testing: Theory and Practice. St. Paul (MN): Assessment Systems Corporation.
23. Wainer, H., Dorans, N. J., Flaugher, R., Green, B.F., Mislevy, R. J., Steinberg, L., & Thissen, D. (1990). Computerized adaptive Testing: A primer. Hillsdale NJ: Erlbaum.
24. Warm, T. A. (1989). Weighted likelihood estimation of ability in the item response theory. Psychometrika, 54, 427-450.
25. Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. Applied Psychological Measurement, 6, 473-492.