Psychological Test and Assessment Modeling, Volume 54, 2012 (1), 73-89

Applying a construction rational to a rule based designed questionnaire using the Rasch model and LLTM

Manuel Reif[1]

Abstract

The use of questionnaires is widespread in psychological assessment. Typically items are constructed more or less "intuitively" and their difficulty is determined with empirical studies. In order to improve the construct validity of questionnaire items an approach to constructing questionnaires by using construction rationals is demonstrated. This approach has its origins in the cognitive sciences and intelligence research and is applied in the present study for the purpose of constructing and validating a self-reporting extraversion scale. Therefore, a construction rational was developed as a basis for item generation allowing the prediction of the difficulty of an item. After establishing a Rasch model fitting item pool, the appropriateness of the developed construction rational was assessed by means of the linear logistic test model (LLTM). It was not possible to fully explain item difficulty with the proposed construction rational. Nevertheless, this approach is a reasonable method for constructing questionnaire items in a more rational rather than intuitive manner. A further benefit to be considered is that an appropriate construction rational would enable automatized item generation.

Key words: IRT, LLTM, Construction Rational, Extraversion, Questionnaire

[1] *Correspondence concerning this article should be addressed to:* Manuel Reif, PhD, Division of Psychological Assessment and Applied Psychometrics, Faculty of Psychology, University of Vienna, Liebiggasse 5, 1010 Vienna, Austria, Europe; email: manuel.reif@univie.ac.at

1 Introduction

Questionnaires are one of the most commonly used methods in psychological assessment when it comes to measuring personality traits. The reason for the popularity of questionnaires is stated as being their economic administration and scoring.

The construction of a self-report personality inventory starts with the decision regarding which trait should be measured. Subsequently, an item pool for the inventory has to be developed. This step is certainly the most crucial point because the items determine the quality of the questionnaire. In order to develop the items, the item-wording has to be carried out in a very thorough manner because even small modifications of the formulation can have a distinct impact on the item properties. For example, it has been found that the use of negatively worded items in addition to positively worded items changes the factor structure (DiStefano & Motl, 2006). However, there are certain recommendations for item generation in the case of self-report surveys (cf. Moosbrugger & Kelava, 2007) – for example, regarding the use of negatively worded items. In comparison to the development of intelligence scales, the development of questionnaire items is commonly carried out in a rather intuitive and heuristic manner, as Janke (1973, cited from Bühner, 2006) remarked. He criticizes items pertaining to the same questionnaire as being too heterogeneous, not specific enough and aimed at totally different aspects of human experiences like interests, behaviors, preferences, opinions etc. These problems are still current.

In the construction of intelligence and achievement tests a rule-based approach exists using a construction rational as a starting point for item generation. This approach originates from cognitive sciences and is related to the statistical techniques of item response theory (IRT), in particular the linear logistic test model (LLTM) by Fischer (1973). To create a construction rational, say for some kind of intelligence scale, means that as a first step the cognitive operations hypothesized as being necessary to solve prototypical items need to be defined. The second step involves combining these cognitive operations in different ways which leads to additional new items. The difficulties of the involved cognitive operations for solving an item add up to the total item difficulty. This approach has already been implemented for various intelligence and achievement tests (e.g. Gittler, 1990; Formann & Piswanger, 1979; Holling, Blank, Kuchenbäcker & Kuhn, 2008). It is also possible, although not recommended, that a construction rational is constructed and applied after the item generation process (e.g. Sonnleitner, 2008; Poinstingl, 2009).

To begin with, the main advantage of the rule based item construction approach (if the construction rational holds the data) is the guarantee of construct validity (cf. Embretson, 2008; Hornke & Rettig, 1989). Constructing items in this way involves testing the underlying psychological theory. If the theory which represents the foundation for the construction rational is not suitable, then a poor model fit for the item pool would result. Additionally, the foundation on a very specific theory determining different cognitive operations which are necessary for solving an item, makes the interpretation of the resulting test score well-founded (Bejar, 2010). A further advantage of this approach is the opportunity to create new items automatically according to the rules of the construction

rational (cf. Arendasy, Sommer, Gittler & Hergovich, 2006). This is especially important in the context of adaptive testing where a large item pool is required (cf. van der Linden, 2008).

As mentioned above, item construction based on item-rules is strongly related to IRT, in particular the LLTM (cf. Embretson, 2008). The LLTM is a special case of the Rasch model which decomposes the item difficulty parameter σ_i from each item into a weighted sum of so-called basic parameters η_j (Fischer, 1974):

$$\sigma_i = \sum_{j=1}^{m} q_{ij}\eta_j \tag{1}$$

The basic parameters describe the properties of the items, and explain their differences, which is why the LLTM is called an "item explanatory model" (De Boeck & Wilson, 2004). The q_{ij} are representing fixed weights, which are specifically defined for each basic parameter in each item.

Several parameter estimation techniques are feasible, but conducting a conditional-maximum-likelihood (CML) estimation guarantees parameter separation and furthers the possibility of "specific objectivity" comparisons (Fischer, 1974) – therefore this method is the most preferable. To check the model fit of the LLTM, the likelihood of the LLTM is compared to the likelihood of the Rasch model by means of a likelihood ratio test (see equation 2).

$$D = -2\log\left(\frac{L_{\mathrm{LLTM}}}{L_{\mathrm{RM}}}\right) \tag{2}$$

As a first step it is essential to prove whether the Rasch model holds the data, because in the further process the Rasch model is supposed to act like a saturated (“true”) model. The second step contains the estimation of the LLTM and the test against the Rasch model, as shown in equation 2. If the LLTM fits the data as well as the Rasch model, equation 2 is approximately $\chi 2$-distributed with p-m, degrees of freedom (whereby p is the number of item parameters and m is the number of the basic parameters η_j). With no significant difference in the likelihood ratio test, the supposed structure (e.g. a construction rational) can be assumed as approved. Furthermore, a non significant result is confirmation of the "construct-validity".

It is difficult for data from intelligence tests to fulfill the assumption that the Rasch model holds. It is even more difficult for questionnaires, and thus the Rasch model is hardly applied to questionnaire data (e.g. Chernyshenko, Stark, Chan, Drasgow & Williams, 2001). One reason is that self-report inventories can be biased by response sets or the tendency of the examinee to answer the items in a “socially desirable” manner. If these tendencies occur, an examinee’s response to an item depends on more than one latent trait – namely the measured personality trait and the response set. Therefore the assumption of unidimensionality which is crucial for the Rasch model is not fulfilled anymore. However, there are a few personality questionnaires which have been evaluated by means, or constructed according to the Rasch model such as the "Trierer Integriertes

Persönlichkeitsinventar (TIPI)" (Trier integrated personality inventory) (Becker, 2002, 2003) and the "Big five plus one" (B5PO) (Holocher-Ertl, Kubinger & Menghin, 2003). The main advantage of constructing a questionnaire by means of the Rasch model is the guarantee that the sum score of a scale is an appropriate measurement. In fact, most of the questionnaires use the sum-scores of each scale to estimate the position on the latent trait continuum of each tested person. But the sum score is only a fair measure if the Rasch model holds the data (Fischer, 1974). Since the LLTM is nothing else but a special restrictive case of the Rasch model, the same benefits are true for it, if it holds for the data set.

As pointed out, applying a construct rational for test construction brings considerable advantages. Therefore it is also desirable that this approach be adopted for the construction of questionnaire items which has been done in the case of the present study.

The first step of the scale construction process is to decide which personality trait shall be measured. For the current study a questionnaire measuring extraversion was constructed. This trait was chosen because it is one of the best examined and most validated traits in psychology. If it is not possible to successfully create a construction rational for this trait it will hardly be possible for any other traits. Furthermore, extraversion questionnaires seem to be less affected by social desirability tendencies (Ferrando, 2008), which is an important point for the model fit of the Rasch model as pointed out above.

Extraversion is part of nearly every big personality theory, such as Eyesenck's biologically founded "Giant-three" (Eysenck, 1944), Jungs "psychological types" (Jung, 1923) and the very widely examined and commonly used "Big Five" as proposed by Costa and McCrae (e.g. Costa & McCrae, 1992). Some well known questionnaires are based upon these theories or slight modifications of them e.g. the "Trier integrated personality inventory" (Becker, 2002, 2003) already mentioned, NEO-PI-R (Costa & McCrae, 1992), the "Eysenck personality profiler" (Eysenck, Wilson & Jackson, 1998), the "Myer Briggs type indicator" (Briggs & Myers, 1991), the "Big five plus one" (Holocher-Ertl et al., 2003) and so on. These questionnaires define extraversion in a similar but not identical manner.

To develop the extraversion questionnaire the definition of extraversion as given in the NEO-PI-R (Costa & McCrae, 1992) was used. In this sense the latent trait extraversion manifests itself as the need for stimulation, the need for interpersonal relationships and the intensity of experiencing joy. After choosing the personality trait for the construction of the questionnaire, the development of the construct rational is the next step in scale construction. Of course, in order to apply the approach of a rule-based item generation to a self-report inventory some modifications are necessary in comparison to the described process in the context of ability and achievement tests. The construction rational for questionnaires is composed of the atomic parts of the items which can be compared to the cognitive operations of an intelligence scale. As an item in a questionnaire cannot be "solved" with certain cognitive operations, these atomic parts have to be defined as a kind of "attraction parameter". The sum of these "attraction parameters" of one certain item should determine the probability of a person agreeing (in the widest sense of the word) with the content of this item. In other words, the "difficulty" of an item is the

result of its basic structure which is determined by the construction rational. In the following, the application of rule-based item generation for the development of an extraversion questionnaire is presented. The appropriateness of the created construct rational is tested by means of the LLTM.

2 Method

As pointed out in the introduction, the definition of extraversion as specified in the NEO-PI-R (Costa & McCrae, 1992) was chosen. In essence, the following facets adjectives represent the basis of the item construction for the questionnaire which was used in this study.

- gregariousness
- positive emotion
- excitement-seeking
- activity
- assertiveness

The items for the extraversion questionnaire were produced in a multilevel process. First of all, the construction rational (shown in figure 1) was created. In the next step, the five facets listed above were presented to a group of experts on psychological assessment. They produced specific, prototypical and at best empirically observable behaviors for each of these traits, which should serve as a basis for the item pool. This approach is called the "Act frequency approach" (Buss & Craik, 1983). The aim of this approach is to infer the latent trait expression from the frequency of a prototypical behavior of a person. In other words, the theory postulates that persons with a high degree of one trait often show more prototypical behavior than an "average" person. Thus the method involves asking examinees about the frequency of their behavior in certain situations. The advantage of this approach is that examinees are not forced to assess themselves with adjectives, vague descriptions or psychological constructs (e.g. "I am dominant."). Subsequently, the different behaviors produced by the experts were sorted with respect to their quality and then categorized into the existing construction rational by a second group of experts on psychological assessment. Next, this collection of categorized behaviors was used to create items consisting of whole sentences according to all the classes of the construction rational. Overall, much attention was paid to optimizing every item according to the item classification proposed by Podsakoff and MacKenzie (2003). Finally, 122 items were ready for administration.

To illustrate the item generation and behavior-centered items based on these schemes, two item examples will be given:

"In ? of 5 cases I start a conversation with an unfamiliar person while shopping." [2]

Shopping is an everyday activity – and so the main purpose of this situation is "daily routine". The stimulative nature of this situation is "neutral" in the context of extraversion. People do not normally seek out shopping to satisfy extraversion-needs. The second person of interaction is unknown – thus the attribute "level of awareness" is "unfamiliar" in this case. In the category "activity" the acting person is "active". The person starts the conversation and begins the extraverted activity intentionally. The "type of communication" is "verbal" – more precisely the "communication style" is categorized as "neutral conversation". This item does not contain any information regarding emotional reactions.

In ? of 5 cases I participate in serious discussions with familiar persons.

The purpose of the situation is "interaction", because the persons involved in the discussion are present. For the same reason, the situation is classified as being of a "highly stimulative nature". Persons who are in such situations most likely want to satisfy extraversion needs. Obviously the persons are classified as "familiar". The acting person merely participates in a discussion but does not initiate it – thus the situation is classified as "passive – participating" because the person joins the discussion but is not the one who starts it. The "type of communication" is "verbal" and the "communication style" is classified as "discussion".

The administration of the newly developed questionnaire was web-based. It was announced in different closed internet forums. Additionally, the questionnaire was administered to psychology students within the framework of a university course, again in a web-based manner. Answering all 122 items conscientiously would have required a lot of effort from an examinee, thus six different booklets were constructed. The linking design of these booklets is shown in figure 2. Every booklet contains about 50% of the total number of items. In order to make answering the questionnaire more varied for the examinees, some items measuring "consciousness" were inserted (these items were also developed according to a construction rational). These consciousness-items will be analyzed in a further study.

To avoid item-position effects (e.g. Ortner, 2004) items were administered in a random order to the examinees.

Overall 560 persons answered the questionnaire. 14 persons had to be excluded from this sample because their reaction times were unrealistically short – thus it must be assumed that they did not answer the questionnaire in a serious manner. The most important sociodemographic characteristics of the remaining 546 examinees can be seen in figure 3. Furthermore the majority (about 88%) of the sample had at least high school graduation.

In order to apply the Rasch model, the answers were dichotomized at the middle of the scale, which means that reactions in categories 0 to 2 were merged to 0, and 3 to 5 merged to 1.

[2] Both items were translated by the author of this article.

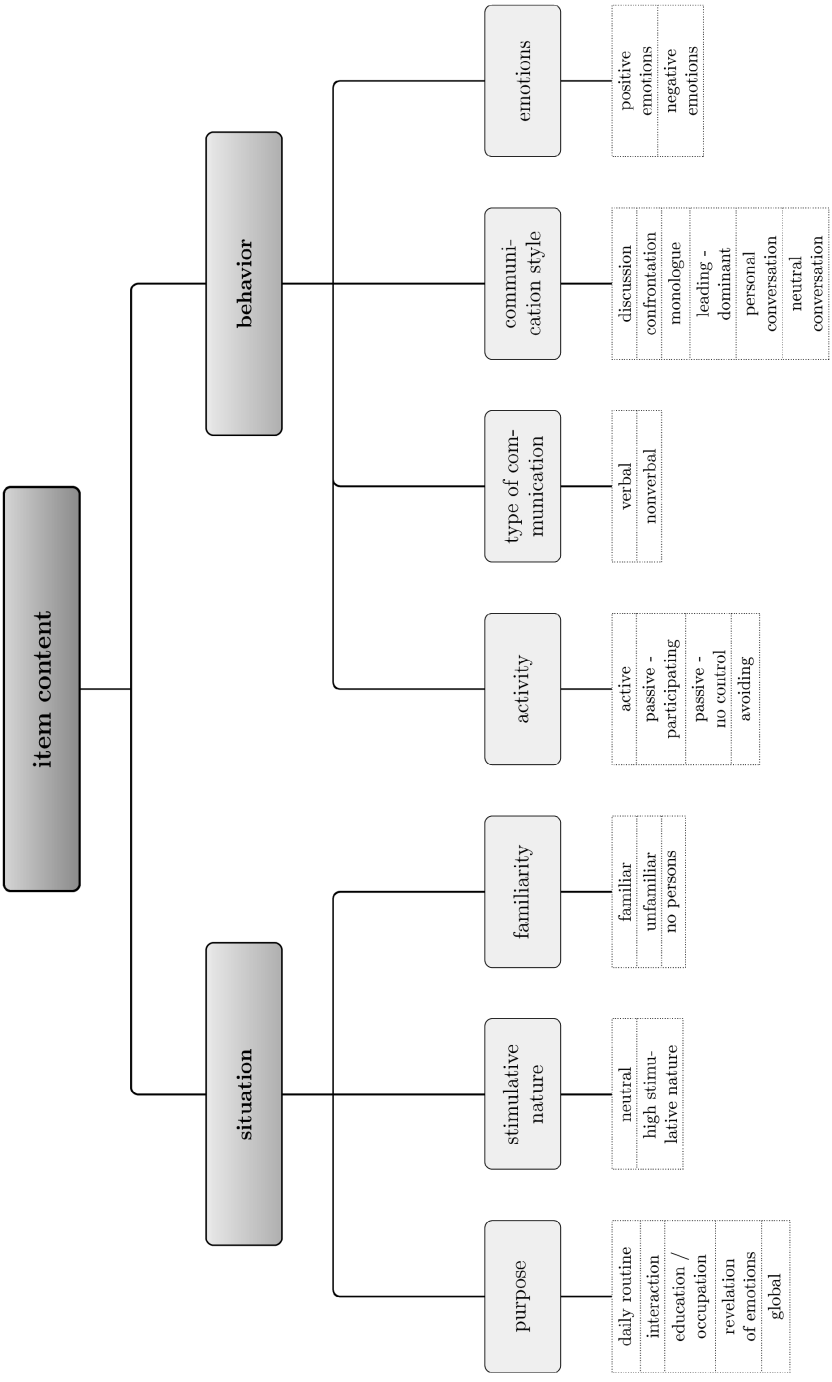

Figure 1: Construction rational

block / booklet	E1	E2	E3	E4	E5	E6	C1	C2	C3
1	■	■	■				■	■	■
2				■	■	■		■	■
3		■		■		■	■		■
4		■	■	■			■	■	
5	■				■	■		■	■
6	■		■		■		■		■

Figure 2:
The six booklets

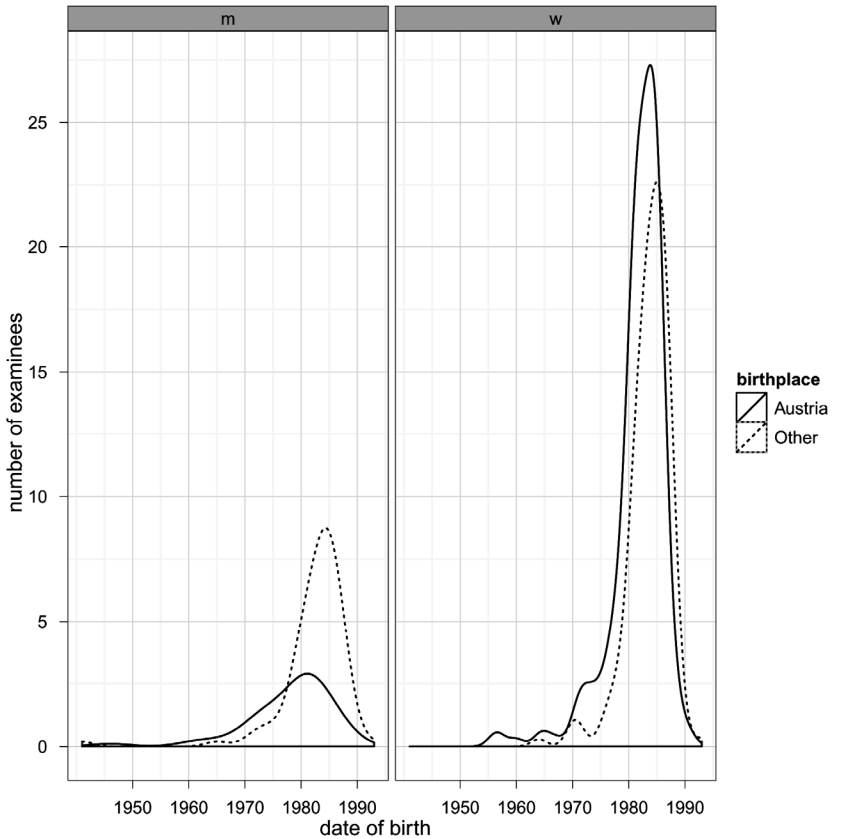

Figure 3:
Sociodemographic characteristics of the sample

3 Results

The data were analyzed according to the Rasch model using eRm (Mair & Hatzinger, 007; see also Poinstingl, Hatzinger & Mair, 2007). To check the model fit, Andersen`s likelihood-ratio test and graphical model checks were conducted, as proposed by Kubinger (2005). Five partition criteria were chosen: low score vs. high score, male vs. female, born in Austria vs. not born in Austria, birth-year lower than 1982 vs. birth-year higher than 1982, education level lower than high school vs. education level equal or higher than high school. The results of the first run of Andersen`s likelihood ratio test, including all 122 items, is shown in table 1. Three partition criteria were statistically significant ($\alpha = 0.01$).

Items with a poor model fit according to the graphical model check and the Wald-Test were deleted step by step. After deleting 18 Items an a-posteriori model fitting item pool was obtained. Table 2 shows the results of Andersen`s likelihood ratio test after excluding 18 non-fitting items.

The misfitting items were examined in more detail to find reasons for the poor fit. It was found that 15 out of the 18 items could be assigned to the category "type of communication – nonverbal" – considering that overall only 35 out of 122 items are assigned to this category. The amount of "nonverbal" items that didn´t fit the model is intriguingly high. But the results are not conclusively unambiguous because the majority of the "nonverbal" items remain as fitting items in the final item pool. If there is a part of extraversion which

Table 1:
Results of Andersen´s Likelihood ratio test – first run

	partition criterion	Andersen χ^2	df	$\chi^2_{\alpha 1\%}$	
1	score	227.897	121	160.1	sig.
2	sex	212.503	121	160.1	sig.
3	age	148.363	121	160.1	not sig.
4	education	124.416	118	156.65	not sig.
5	birthplace	201.759	121	160.1	sig.

Table 2:
Results of Andersen´s Likelihood ratio test after deleting 18 items

	partition criterion	Andersen χ^2	df	$\chi^2_{\alpha 1\%}$	
1	score	136.471	103	139.3	not sig.
2	sex	119.632	103	139.3	not sig.
3	age	129.144	103	139.3	not sig.
4	education	100.309	100	135.8	not sig.
5	birthplace	133.365	103	139.3	not sig.

must be considered as a unique and independent factor of "verbal" extraversion-behavior then this could be a topic for further studies.

The LLTM was applied for the remaining 104 fitting items, again using eRm. The weight matrix Q was designed by means of the construction rational. The basic parameters represent the item properties as pictured in figure 1. The parameters (categories) an item is assigned to are set to 1 – the other parameters (categories) are set to 0. The matrix Q was always checked for full rank – redundant columns were deleted. The first few lines of the first matrix Q are shown in table 3.

To check if the LLTM holds against the Rasch model, a likelihood ratio test (according to formula 2) and a graphical model check were carried out. The likelihood ratio test resulted in significance as displayed in table 4. The graphical model check (see figure 4) shows strong deviations from the postulated line of 45 degrees. It is obvious that these 18 basic parameters cannot explain the item parameters of the Rasch model. In particular, the lowest and highest parameters according to the Rasch model are estimated less extremely.

Therefore, three additional components were added to the construction rational. These three new components refer to the structural design of the items rather than to the item content.

Negatively worded items About a third of the items are negatively worded, and had to be scored in the reverse direction. The potential impact on the answering process was taken into consideration by introducing an additional parameter for these items.

Decision question About 10% of the items incorporate specific behaviors, which are presented in contrast to an alternative behavior. So the examinee specifies the frequency of one behavior in relation to the other – in other words the examinee has to decide which behavior is shown more often in the situation mentioned by the item.

Table 3:
First 5 lines of matrix Q

	afi	afb	afg	gg	afa	exg	ub	ak	pt	ve	nv	di	kf	mo	fd	pg	poe	ne
1	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0
2	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0
3	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0
4	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0
5	0	0	1	0	0	1	1	1	0	0	0	0	0	0	0	1	0	0
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:

Table 4:
Likelihood-ratio test: Rasch model vs. LLTM-1

$logL_{lltm}$	$logL_{rm}$	$-2(log(L_{lltm})-log(L_{rm}))$	$Chi^2_{1\%}$	df	
-14421.51	-12970.38	2902.38	118.24	85	Sig.

Frequency of situation Certain situations occur more often than other situations. For instance, "going shopping" will occur, in most cases, more often in an average person`s life than "going to the hairdresser". Thus the items were classified as to whether the situations occur "very often", "on an average basis" or "rarely". The reason for this classification is the fact that situations that rarely occur are supposedly not as easy to remember as situations that occur on a daily basis.

These parameters extended the weight matrix Q by four parameters. Again the LLTM was applied, now with 22 basic parameters. The result of the likelihood ratio test for this “extended construction rational” is shown in table 5. Compared to the first model the improvement of fit is obvious but not great and nevertheless the likelihood ratio test is highly significant. There is hardly any improvement in the graphical model check (see figure 5) when compared to the previous one. Obviously, these new parameters did not have such a great effect on the model fit.

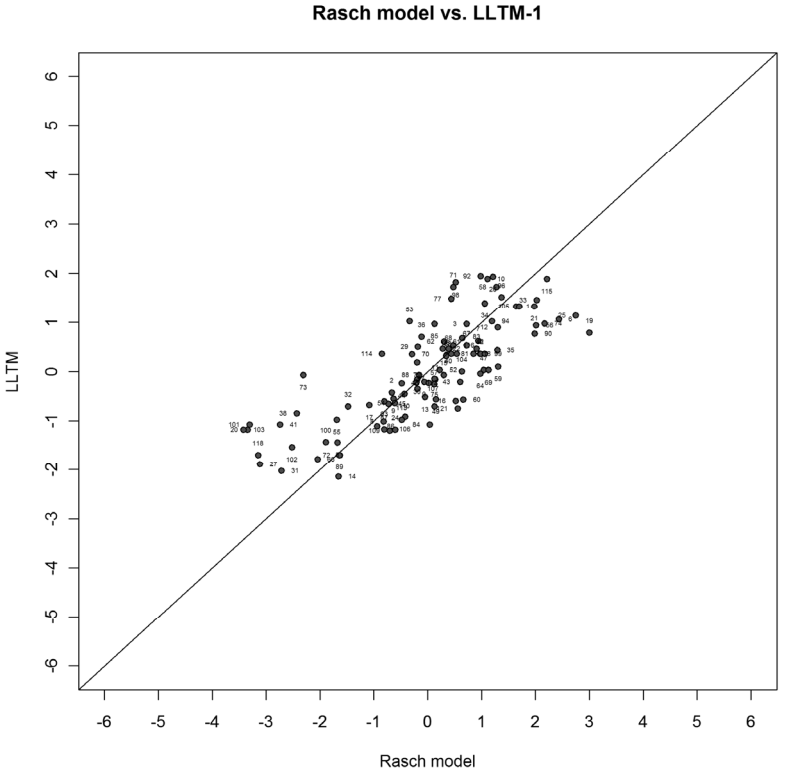

Figure 4:
Rasch model vs. LLTM-1

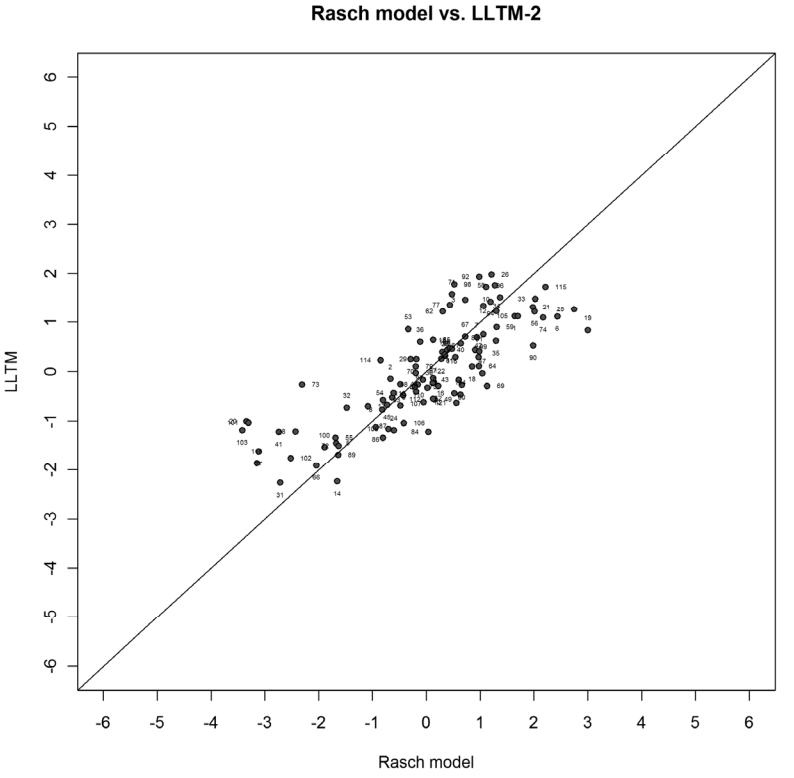

Figure 5:
Rasch model vs. LLTM 2

Table 5:
Likelihood-ratio test: Rasch model vs. LLTM-2

$logL_{lltm}$	$logL_{rm}$	$-2(log(L_{lltm})-log(L_{rm}))$	$Chi^2_{1\%}$	df	
-14314.16	-12970.38	2687.55	113.51	81	sig.

In order to find out what kind of additional component for the construction rational is needed to improve model fit, difficulty parameter estimates from the Rasch model and LLTM were compared. Items whose estimates differ by more than 1.5 scale units were investigated. Six items were consequently flagged as suspicious. In five cases, the item difficulty was underestimated by the LLTM. These five items are somehow similar – the situations displayed are of an extremely low stimulative nature. Thus, an additional component was defined for the category "main purpose of the situation" which refers to situations with an extremely low stimulative nature.

Unfortunately, the introduction of four new parameters into the second Q matrix only lead to a slight improvement of item fit, which could not justify the increase of the number of parameters. For this reason a parsimonious model was defined by leaving out three of the four parameters. Only one parameter, "decision question", remained in the third matrix Q because this parameter was, compared to the other new parameters, the most different from zero. Thus, the third matrix Q contains all the parameters of the original construction rational, "decision question" and, as mentioned before, the new parameter.

The third time that the LLTM and the likelihood ratio test were carried out there was again a significant result as shown in table 6. Even the second enhancement of the construction rational did not lead to a satisfying result. The graphical model check, shown in figure 6, shows a distinct improvement – but nevertheless noticeable deviations. A com-

Table 6:
Likelihood-ratio test: Rasch model vs. LLTM-3

$logL_{lltm}$	$logL_{rm}$	$-2(log(L_{lltm})-log(L_{rm}))$	$Chi^2_{1\%}$	df	
-14010.9	-12970.38	2081.03	115.88	83	sig.

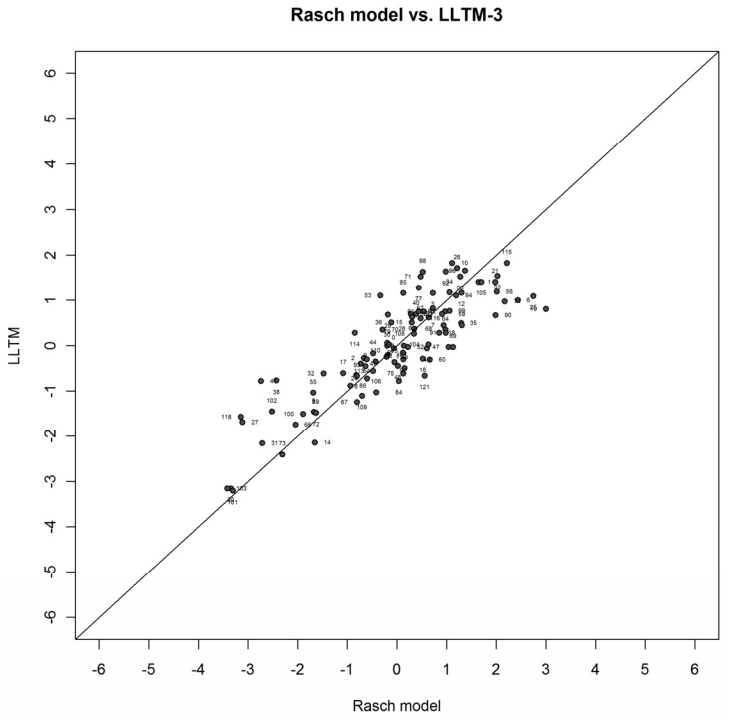

Figure 6:
Rasch model vs. LLTM 3“

parison of the three LLTM estimations is shown in table 7. The result is not surprising – the third model shows the best model fit. A next possible step could be to delete non fitting items according to the graphical model check, in order to adjust the data to the model. But this approach was not pursued because the main idea of this study was to validate the item pool with its construction rational. Deleting items would have meant that the construction rational is not correct, as already known.

Unfortunately the construction rational cannot explain the data as expected. Despite the construction rational fail to fit the data, it can at least be used to roughly predict item difficulty when creating new items.

Table 7:
The three models compared with information indices

	AIC	BIC
LLT M_1	28879.02	29069.91
LLT M_2	28672.31	28905.63
LLT M_3	28061.80	28273.90

4 Discussion

In the present study an approach to test construction which originates in the cognitive sciences is applied in order to develop a self-report questionnaire measuring extraversion. Items were generated strictly according to a priori determined construction rules and an a priori developed construction rational.

Firstly, the model fit of the Rasch model was assessed revealing a rather good fit. Only 18 of 122 items had to be excluded due to a poor item fit. Thus, it can be said that this first step at least resulted in an item pool calibrated to the Rasch model which offers the possibility of adaptive testing.

Subsequently, the appropriateness of the developed construction rational was assessed by means of the LLTM. Results indicate that the proposed construction rational cannot explain item difficulties sufficiently. After the inspection of items with the most extreme deviations the construction rational was modified twice. These modifications improved the model fit but nevertheless resulted in significant results. Therefore, it has to be concluded that the postulated construction rational is not appropriate in order to explain the item difficulties. Obviously there are some facets influencing the difficulty of items which were not modeled. One hypotheses is that certain combinations of categories of the construction rational result in a special item quality. In other words, some of the difficulty parameters could be considered to be more than the addition of the appropriate basic parameters. It seems necessary to revise the postulated construction rational for the existing questionnaire. Adding new categories would maybe require new items to be created of these classes as well. In any case, a new data collection process would then be needed to evaluate this severely modified construction rational.

Another critical point of the study is that the data had to be dichotomized before analysis due to the sample size. Adding categories together is always a critical point, which could also have contributed to the poor model fit. A further study providing data of a much larger sample size could assess the fit of a linear partial credit model. This model takes typical questionnaire data with more than two response categories into account but a larger sample size than in the present study is necessary to estimate this model.

Although the developed construction rational turned out to be unsuitable, the described approach should be considered as a new and methodically well founded method for further personality scale constructions. The advantages are obvious: first of all, this approach improves construct validity and offers the possibility of automatic item generation. Even in the present case, being that the construction rational does not include every facet that affects item difficulty, the prediction of the difficulty of newly created items would be better than only an intuitively given prediction without empirical foundations.

In summary, it can be said that creating a construction rational for questionnaires is not as straightforward as for intelligence-scale items – however, it is still necessary to attempt to make questionnaires more valid and reliable.

References

Arendasy, M., Sommer, M., Gittler, G., & Hergovich, A. (2006). Automatic generation of quantitative reasoning items. *Journal of Individual Differences, 27*, 2-14.

Becker, P. (2002). Das Trierer Integrierte Persönlichkeitsinventar [Trier integrated personality inventory]. *Diagnostica*, *48* (2), 68-79.

Becker, P. (2003). *Trierer Integriertes Persönlichkeitsinventar (TIPI) [Trier integrated personality inventory (TIPI)]*. Göttingen: Hogrefe.

Bejar, I. I. (2010). Recent Development and Prospects in Item Generation. In Embretson, S. E. (Eds.), *Measuring Psychological Constructs. Advances in Model-Based Approaches* (pp. 201-226). Washington: American Psychological Association.

Bühner, M. (2006). *Einführung in die Test- und Fragebogenkonstruktion. [Introduction to test and scale construction]. (2nd ed.)* München: Pearson

Buss, D. M., & Craik, K. H. (1983, April). The Act Frequency Approach to Personality. *Psychological Review*, *90* (2), 105-126.

Chernyshenko, O. S., Stark, S., Chan, K.-Y., Drasgow, F., & Williams, B. (2001). Fitting Item Response Theory Models to Two Personality Inventories: Issues and Insights. *Multivariate Behavioral Research*, *36* (4), 523-562.

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO personality inventory (NEO-PI- R) and NEO Five-Factor Inventory professional manual.* Odessa: Psychological Assessment Resources.

De Boeck, P., & Wilson, M. (2004). *Explanatory Items Response Models*. New York: Springer.

DiStefano, C., & Motl, R. W. (2006). Further Investigating Method Effects Associated With Negatively Worded Items on Self-Report Surveys, *Structural Equation Modeling, 13,* 440-464.

Embretson, S. E., & Daniel, R. C. (2008). Understanding and quantifying cognitive complexity level in mathematical problem solving items. *Psychology Science Quarterly, 50,* 328-344.

Eysenck, H. (1944). Types of personality – a factorial study of 700 neurotics. *Journal of Mental Science*, *90* , 851-861.

Eysenck, H., Wilson, G., & Jackson, C. (1998). *Eysenck Personality Profiler (EPP–D)*. Frankfurt: Swets Test Services.

Ferrando, P. J. (2008). The impact of social desirability bias on the EPQ-R item score: An item response theory analysis. *Personality and Individual Differences, 44,* 1784-1794.

Fischer, G. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37* , 359-374.

Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests [Introduction to the theory of psychological tests]*. Bern: Huber.

Formann, A., & Piswanger, K. (1979). *Wiener Matrizen-Test (WMT) [Viennese Matrices]*. Weinheim: Beltz.

Gittler, G. (1990). *Dreidimensionaler Würfeltest (3DW)[Three-dimensional Cube test]*. Hogrefe.

Holling, H., Blank, H., Kuchenbäcker, K., & Kuhn, J.-T. (2008). Rule-based item design of statistical word problems: A review and first implementation. *Psychology Science Quarterly, 50*, 363-378.

Holocher-Ertl, S., Kubinger, K. D., & Menghin, S. (2003). Big Five Plus One Persönlichkeitsinventar (B5PO) [Big Five Plus One Personality Inventory (B5PO)] [Software und Manual]. Mödling: Dr. G. Schuhfried GmbH.

Hornke, L. F., & Rettig, K. (1989). Regelgeleitete Itemkonstruktion unter Zuhilfenahme kognitionspsychologischer Überlegungen [Rule-based item generation using cognitive psychology]. In K. D. Kubinger (Hrsg.), *Moderne Testtheorie [Modern psychometrics]* (S. 140-162). Weinheim: Beltz.

Janke, W. (1973). Das Dilemma von Persönlichkeitsfragebogen. Einleitung des Symposiums über Konstruktion von Fragebogen [The dilemma of personality questionnaires. Introduction of the symposium on the construction of questionnaires]. In G. Reinert (Hrsg.), *Bericht über den 27. Kongress der Deutschen Gesellschaft für Psychologie in Kiel 1970 [Report from the 27th Congress of the German Society of Psychology in Kiel 1970]* (S. 44-48). Göttingen: Hogrefe.

Jung, C. (1923). *Psychological types*. New York: Harcourt Brace.

Kubinger, K. D. (2005). Psychological Test Calibration Using the Rasch Model – Some Critical Suggestions on Traditional Approaches. *International Journal of Testing*, *5* (4), 377-394.

Mair, P., & Hatzinger, R. (2007). eRm: Extended Rasch modeling. R package Vs 0.9.5. http://cran.r-project.org/

Moosbrugger, H., & Kelava, A. (Eds). (2007). *Testtheorie und Fragebogenkonstruktion [Test theory and Scale construction]*.Heidelberg: Springer.

Ortner, T. (2004). On changing the position of items in personality questionnaires. Analysing effects of item sequence using IRT. *Psychology Science Quaterly*, *46* (4), 466-476.

Podsakoff, P. M., & MacKenzie, S. B. (2003). Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies. *Journal of Applied Psychology*, *88* , 879-903.

Poinstingl, H., Mair, P., & Hatzinger, R. (2007). *Manual zum Softwarepackage eRm (extended Rasch modeling). Anwendung des Rasch-Modells (1-PL Modell) – Deutsche Version* [Manual of eRm.To apply the Rasch model – German Version]. Lengerich: Pabst.

Poinstingl, H. (2009). The Linear Logistic Test Model (LLTM) as the methodological foundation of item generating rules for a new verbal reasoning test. *Psychology Science Quaterly*, *51* (2), 123-134.

Sonnleitner, P. (2008). Using the LLTM to evaluate an item-generating system for reading comprehension. *Psychology Science Quaterly*, *50* (3), 345-362.

Van der Linden, W. J. (2008). Some New Developments in Adaptive Testing Technology. *Journal of Psychology, 216,* 3-11.