

Mixed Rasch Models for Analyzing the Stability of Response Styles Across Time: An Illustration with the Beck Depression Inventory (BDI-II)

Ferdinand Keller and Ingrid Koller

Abstract Questionnaires for clinical studies are often evaluated in cross-sectional settings and on the basis of classical test theory. Some of them, like the BDI-II which is one of the most widely used self-report instruments for assessing depression severity, are considered to have very good psychometric properties. However, these properties are rarely evaluated in longitudinal designs, and even less with models of item response theory (IRT). In addition, analyses of self-report questionnaires with IRT models provided evidence of two major response styles: the tendency to prefer extreme response categories, and the tendency to prefer the middle categories. Rasch models, in particular their extension to the so-called mixed Rasch model, are well suited to address these questions. They allow one to determine latent classes with different response styles and to analyze qualitative aspects of change such as the consistency of response styles across time. In this chapter first, an introduction to response styles and an overview of the mixed Rasch model, especially in the context of measuring change, are given and second, a practical example is elaborated using a sample of in-patients from a psychosomatic clinic that were assessed with the BDI-II at the beginning and at the end of in-patient treatment. The presence of two response styles is confirmed for the admission data, whereas for the discharge data the Rasch model seems sufficient. A combined analysis of both time points reveals three classes, one of which is a low symptom class and the other two reflect, again, the two response styles; these two classes remain quite stable over time.

F. Keller (✉)

Department of Child and Adolescent Psychiatry and Psychotherapy,
University Hospital of Ulm, Steinhoevelstr. 5, 89075 Ulm, Germany
e-mail: ferdinand.keller@uniklinik-ulm.de

I. Koller

Department for Psychology, Alpen-Adria-Universität Klagenfurt,
Klagenfurt, Austria

Measurement of Change in Clinical Psychology and Response Styles

Measurement of change in clinical psychology and psychiatry is of major importance for the evaluation of treatment approaches that are suited best for patient groups (i.e., comparing different types of psychotherapy and/or psychopharmacological treatment) as well as for monitoring improvement on an individual level (e.g., is there clinically significant progress across the treatment sessions, or is it indicated to modify the treatment approach?).

Unlike in achievement research (e.g., measurement of educational trajectories) where sophisticated statistical models are applied to assess the psychometric properties of items and to investigate change across time, treatment evaluation studies in the clinical realm mostly rely on a few, well-established outcome instruments that have high clinical face validity but whose psychometric properties in designs with repeated measurement are nonetheless rarely tested. Although this facilitates the comparison of study results, the measurement properties in longitudinal designs are largely unknown, except the test-retest-reliability based on the classical test theory approach.

A further threat for reliability and validity of self-report measures using Likert-type response scales are response styles which denote the tendency of an individual to respond to items irrespective of content. Plieninger and Meiser (2014) and Wetzel, Carstensen, and Böhnke (2013) give an overview on research regarding different types, in particular the extreme response style (ERS), i.e., the tendency to prefer the extreme response categories, and midpoint responding (MRS), i.e., the tendency to choose the middle categories (other response styles are, e.g., acquiescence and its opposite, disacquiescence). The authors conclude that past research suggested that response styles may be conceptualized as trait-like constructs that are stable across content domains and time. However, Weijters, Geuens, and Schillewaert (2010) question the results of previous studies on stability over time because of several methodological problems that arise with longitudinal designs, in particular possible memory effects and the usage of the same items which makes it impossible to distinguish between common variance due to response style and due to content.

In this chapter, we focus on the assessment of depression with a self-report instrument, the Beck Depression Inventory in its revised version (BDI-II; Beck, Steer, & Brown 1996; German version: Hautzinger, Keller, & Kühner 2006). The BDI-II is one of the most widely used self-report instruments to assess severity of depression in treatment studies as well as in psychodiagnostics. The psychometric properties are considered to be very good and extensive factor analytic studies have been done on cross-sectional samples (e.g., Brouwer, Meijer, & Zevalkink 2013a; Bühler, Keller, & Läge 2014; Ward 2006).

The BDI-II is used to address the presence of ERS and MRS in a clinical context. Moreover, the stability or change of these (potential) response styles across two time points (admission and discharge in a psychosomatic hospital) and the impact on the measurement of depression severity are examined. To our knowledge, neither issue

has been addressed before in the literature. Furthermore, the assessment of stability is confounded by the clinical intervention (treatment of the patients during their hospital stay) and thus more complicated than in studies where relatively stable traits (personality or achievement) are analyzed. Relations to basic variables which are available for this sample (gender, age, as well as diagnostic subgroups) will be assessed, too. Our method of choice is the mixed Rasch model (MRM; Rost 1990; Rost & von Davier 1995) which is an item response theory model (IRT) that is well suited to identify subgroups of patients that differ in response style, and offers the possibility to assess qualitative change across time (e.g., Glück & Spiel 1997). In the next sections the MRM and its application in the context of assessing different response styles and measuring change are described. After that the empirical example with the BDI-II is elaborated using the MRM approach.

The Mixed Rasch Model

The MRM is a generalization of the Rasch model (RM; Rasch 1960) to a discrete mixture distribution model which makes it possible to extract latent classes of individuals within which the RM holds. Between the extracted classes the RM has not to fit the data and, therefore the order of item difficulty and the range of item difficulties are allowed to vary. Thus, different response scale category usage can exist and therefore RM properties, e.g., measurement invariance, are not given between latent classes (e.g., Baghaei & Carstensen 2013; Embretson 2010; Meiser, Hein-Eggers, Rompe, & Rudinger 1995; Rost, Carstensen, & von Davier 1999; Rost & von Davier 1995). In summary, the MRM combines the unidimensional Rasch model with latent class analysis (LCA; e.g., Meiser et al. 1995; Meiser 2010; Rost 1991). But contrary to LCA, where within classes no person ability variation is assumed, MRM allows the quantification within classes, which means that individuals can differ in ability (e.g., Rost 2004; Spiel & Glück 2008).

In addition to the MRM for two-categorical items, extensions for items with polytomous response formats exist, for example, the mixed partial credit model (PCM) and the mixed rating scale model (RSM; e.g., Von Davier & Rost 1995). Because the applied example in this chapter is based on a polytomous response format the equation for the mixed PCM and one restriction, the mixed RSM, are shown. The restriction to the MRM is straightforward and is explained in Rost and von Davier (1995).

The mixed PCM defines the probability for a person $v = 1, \dots, n$ to pass the threshold $l = 1, \dots, m$ (with $s = 0, \dots, m$ categories) of an item $i = 1, \dots, k$ given the person ability θ_v in class $c = 1, \dots, C$ and the item difficulty β_{ilc} with

$$P(x_{vile} = l | \theta_{vc}, \beta_{ilc}) = \sum_{c=1}^C \pi_c \frac{\exp(l\theta_{vc} - \beta_{ilc})}{\sum_{s=0}^m \exp(s\theta_{vc} - \beta_{isc})},$$

where π_c is the probability to belonging in latent class c (class size parameter) and the item difficulty $\beta_{ilc} = \sum_{l=1}^m \tau_{ilc}$, with the normalization $\sum_{i=1}^k \sum_{l=1}^m \tau_{ilc} = 0$, and $\beta_{i0c} = 0$ within all classes (see also, Rost 1991 or Wetzel et al. 2013). Furthermore, the mixed RSM results from the restriction $\tau_{ilc} = \beta_{ic} + \tau_{sc}$ where the same distances between thresholds are assumed for all items within all classes.

MRM fit will be tested in two ways. First, to test whether the estimated model fits the data, it has to be compared with the saturated model (i.e., the model with the maximum of estimable parameters) by a likelihood ratio test or Pearson chi-square test (see, e.g., Spiel & Glück 2008). Second, the estimated models (e.g., two-class and three-class solution) have to be compared using information criteria, such as, the Akaike information criterion (AIC; Akaike 1974), Bayesian Information Criterion (BIC; Schwarz 1978), or Consistent Akaike Information Criterion (CAIC; Bozdogan 1987). Based on the literature (Baghaei & Carstensen 2013; Wetzel et al. 2013) and simulation studies for the evaluation of performance of information criteria (Preinerstorfer & Formann 2012) BIC and CAIC should be preferred. A qualitative goodness of fit check is the comparison of the average membership probability of different individuals. If it is possible to assign individuals with high probability to one class, the MRM describes the data or response patterns well (see, Spiel & Glück 2008).

Assessment of Response Styles with the MRM

Several studies exist where the MRM was applied to various types of data for the detection of response styles in achievement tests (e.g., Baghaei & Carstensen 2013; Spiel & Glück 2008) and in personality questionnaires (e.g., Eid & Zickar 2010; Gollwitzer, Eid, & Jürgensen 2005; Rost et al. 1999; Rost, Carstensen, & von Davier 1997). All studies showed the suitability of the MRM for the identification and better understanding of different response styles. For example, Wetzel et al. (2013) analyzed several PISA 2006 attitude scales and the subscales of the NEO-PI-R with mixed PCM and further combined the respective latent response classes by means of a second order latent class analysis (c.f. Keller & Kempf 1997). The authors found that for 77 % of the participants a response style (ERS or MRS) occurred consistently across traits.

Furthermore, Wetzel et al. (2013) state that testing the consistency of response styles with the MRM requires that participants only differ in their response style but not in the trait that is being assessed or other factors that might influence the choice of a response category. Thus, the authors recommend estimating a constrained PCM where item locations are fixed to be equal across classes.

Assessment of Response Styles Across Time

In addition to the assessment of response styles in general, it can be of interest to determine whether class membership and, therefore, response style change over time. It is also possible to investigate this kind of question with MRMs (Glück & Spiel 1997 2010; Spiel & Glück 1998). With this exploratory approach it is possible to assess qualitative change across time. Research questions could be whether the class membership is constant over time or whether changes in membership are constant over time (e.g., those associated with class one at time point one change primarily to class two at time point two). For applications of the MRM in the case of dependent data, see, e.g., Glück and Spiel (1997 2010), Meiser et al. (1995), Meiser, Stern, and Langeheine (1998), and Rost (2004).

Technically, the data matrix has to be rearranged before analysis depending on research question. Two examples can be seen in Fig. 1 (see Rost 2004). Further possibilities for longitudinal data are conceivable (see, e.g., Meiser et al. 1998), but not of interest for our study and thus not discussed in this chapter.

If the data matrix is rearranged as shown in the left panel of Fig. 1 (long-format), one gets twice (or t times) as many participants, and change can be analyzed in one step (e.g., Glück & Spiel 1997). Thus, each time point can be seen as an independent subgroup of individuals. The individuals starting from t_2 are called virtual individuals. With this approach the item parameters are estimated in one step

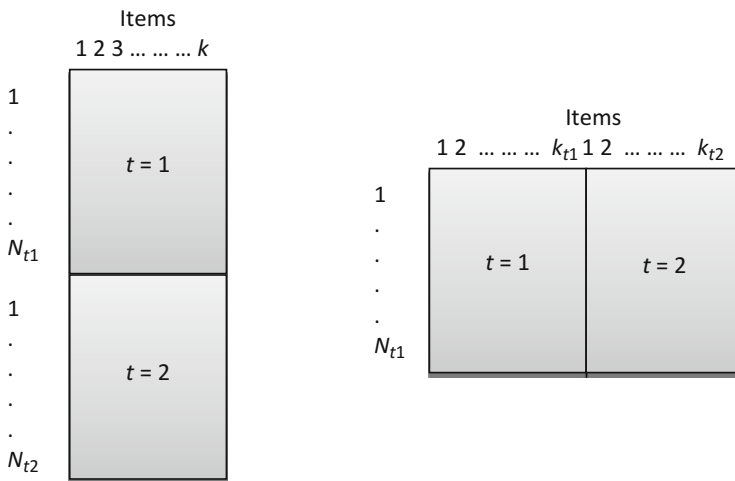


Fig. 1 Two possible ways to rearrange the data matrices for MRM in longitudinal studies. *Left panel* (long format): Data matrix with virtual persons at t_2 . With this rearranging twice (or t) as many persons are available for analysis of dependent data. The MRM analysis can be performed in one step for all time points but the instrument must contain the same items across time. *Right panel* (wide format): Data matrix with virtual items at t_2 . The MRM analysis can be performed also in one step

and it can be seen whether the individuals are staying within or moving between classes. However, there is one restriction, that is, the tests must contain the same items at all time points. In addition it must be taken into account that the assumption of local independence on the person side is violated.

It is also possible to rearrange the data as shown in the right panel in Fig. 1 (wide-format), and to analyze the time points as one long test. Again, in this approach the item parameters are estimated in one step, but, in addition, classes of participants are identified whose items at, e.g., t_2 reflect different magnitudes of change and different types of change (see, Glück & Spiel 1997). This approach, however, hides one major drawback. Due to the prolonged test, the sample size must be increased for a sufficiently accurate estimation of item parameters.

In the previous sections, the application of the MRM when assessing response styles and the procedure for the investigation of qualitative change in dependent data were described. In the next section we show the assessment of response styles using the MRM in a clinical context. First, the sample, the BDI-II and the procedure are described and second the results are given and discussed.

Assessment of Response Styles with the BDI-II

Sample

The sample consisted of in-patients from a clinic for psychosomatic disorders ($N = 1164$); they completed the BDI-II at admission within the routine diagnostic procedure and also at discharge. The mean age in the sample was 45.2 years ($SD = 10.8$; range: 19–72) and 64.7 % of the patients were female. The mean BDI-II total score at admission was 21.4 ($SD = 10.6$) and at discharge 9.1 ($SD = 8.1$). Eight hundred and two patients (68.9 %) were diagnosed with a primary affective disorder (ICD-10: chapter F3) as their main diagnosis; when taking F3 as a comorbid diagnosis, 1001 patients (86.0 %) fulfill the criteria of a depression. The most frequent comorbid disorder was substance-related disorders (ICD-10: F1; $n = 254$ (21.8 %)), and within a range of 15–19 % were somatoform disorders, anxiety-related disorders and post-traumatic stress disorder (PTSD), eating disorders, and personality disorders (see Table 3).

Description of the BDI-II

The BDI-II consists of 21 items that assess a wide range of depressive symptoms (e.g., sadness, suicidal thoughts and wishes, concentration difficulty, or loss of energy). Each item has four categories numbered from 0 to 3 that are formulated in a symptom-specific way (e.g., item 9 “suicidal thoughts and wishes” has the four

response options: 0 = “I don’t have any thoughts of killing myself,” 1 = “I have thoughts of killing myself, but I would not carry them out”, 2 = “I would like to kill myself,” and 3 = “I would kill myself if I had the chance”). The total score of these items reflect the severity of depression. In 1996, a minor revision of the BDI was carried out to meet the criteria of the DSM-IV (American Psychiatric Association 1994) and resulted in the BDI-II (Beck et al. 1996). Symptom scores from 14 to 19 indicate a mild depression, 20 to 28 a moderate, and above 28 a severe depression (Beck et al. 1996).

Procedure

The software program WINMIRA v1.45 (Von Davier 2001) was used to estimate the MRMs. We restricted ourselves for this data example to the mixed PCM, since it has been found in several samples that the fit of the RSM was worse than the fit of the PCM (Keller 2012), which supports the theoretical assumption that the BDI-II with its symptom- and category-specific text requires no restrictions on the category thresholds. The number of latent classes was successively increased from the PCM (1-RM) up to a PCM with three latent classes (3-RM) and parsimony of the models was evaluated using BIC and CAIC, as described above. Participants are then assigned to their most probable class and frequency tables are used to explore relations between time points and to the demographic variables. To compare the identified latent classes and to test the fit of the PCM, MRM analyses are performed, first, for the two time points separately, and then for the virtual sample (long-format, see Fig. 1, left panel) as suggested by Glück and Spiel (1997) and Rost (2004). Additionally, to test the model fit of the final solution (critical $\alpha = 5\%$), 500 re-simulations were carried out and the Pearson χ^2 test-statistic was calculated (see Langeheine, van de Pol, & Pannekoek 1996); according to the recommendation in the WINMIRA output, only the p -value of the empirical probability distribution is reported.

An MRM analysis of the virtual items (wide-format, see Fig. 1, right panel) in one step was omitted, since it runs into several problems: (a) the number of estimated parameters gets in misbalance with our sample size (e.g., for two latent classes almost 500 parameters have to be estimated); (b) the dimensionality of item parameters could be tested, in particular the interesting question whether the items at t_1 and the items at t_2 are homogeneous, but the result would be valid only for this special split of items (t_1 vs. t_2). There is no analogue to the MRM for determining person heterogeneity (where two or more groups (latent classes) are built to achieve maximum person heterogeneity between classes) for the detection of maximum item heterogeneity (Rost 2004).

Following Wetzel et al. (2013), a constrained PCM is also estimated where the item locations are fixed to be equal across classes. The constrained PCM delivers homogeneous latent classes which only differ in the distribution of the threshold parameters (Wetzel et al. 2013) that is in response style. Consequently, the authors

compare the unconstrained PCM with the constrained PCM and use only those subscales for which the constrained PCM (i.e., ensuring trait homogeneity between the latent classes) shows a better fit in BIC and CAIC than the unconstrained PCM.

Results

Mixed PCM Estimated Separately for the Two Time Points

The likelihood, number of parameters, and the information criteria for the PCM and the two-class and the three-class solution are displayed in Table 1. For the admission data, there is a clear minimum in BIC and CAIC for the solution with two latent classes (Modelfit_{2Class}: empirical $p = .046$). The first class consists of 64.3 % of the individuals, and the thresholds (see Fig. 2) suggest that this class prefers to use the middle categories. The estimated thresholds for the second class (35.7 %) are closer together; that is, it is more difficult for them to “leave” category zero and also not very difficult to endorse the highest category: they prefer the extreme categories. Item 9 (suicidal thoughts) has a high threshold in both classes, because acute suicidality is an exclusion criteria in a psychosomatic clinic and thus, the frequencies

Table 1 Model fit for the PCM at admission, at discharge (both estimated separately), and for the virtual sample (long format)

Partial credit models	Log-Lik.	# of parameters	BIC	CAIC	Reliability	Class sizes (%)
<i>Admission</i>						
1-RM	-26,227.83	125	53,338.1	53,463.1	.91	100
2-RM	-25,640.07	249	53,038.0	53,287.0	.91/.92	64/36
3-RM	-25,344.08	373	53,321.4	53,694.4	.89/.92/.90	43/32/25
2-RM constr.	-25,803.55	228	53,216.7	53,444.7	.90/.90	61/39
<i>Discharge</i>						
1-RM	-17,714.72	125	36,311.9	36,436.9	.84	100
2-RM	-17,261.26	249	36,280.4	36,529.4	.73/.87	63/37
2-RM constr.	-17,335.96	228	36,281.5	36,509.5	.75/.87	64/36
<i>Long format</i>						
1-RM	-45,483.88	125	91,936.9	92,061.9	.90	100
2-RM	-44,153.86	249	90,238.2	90,487.2	.90/.78	54/46
3-RM	-43,496.61	373	89,885.0	90,258.0	.77/.90/.91	41/38/21
4-RM ^a						
3-RM constr.	-44,015.01	331	90,596.2	90,927.2	.90/.72/.89	41/33/26

Note. constr. = constrained, i.e., item locations set equal across classes

^aSeveral attempts to estimate a four-class solution resulted always in non-convergent solutions and the fourth class consists of almost no person (class sizes <0.1 %); the other classes remain the same

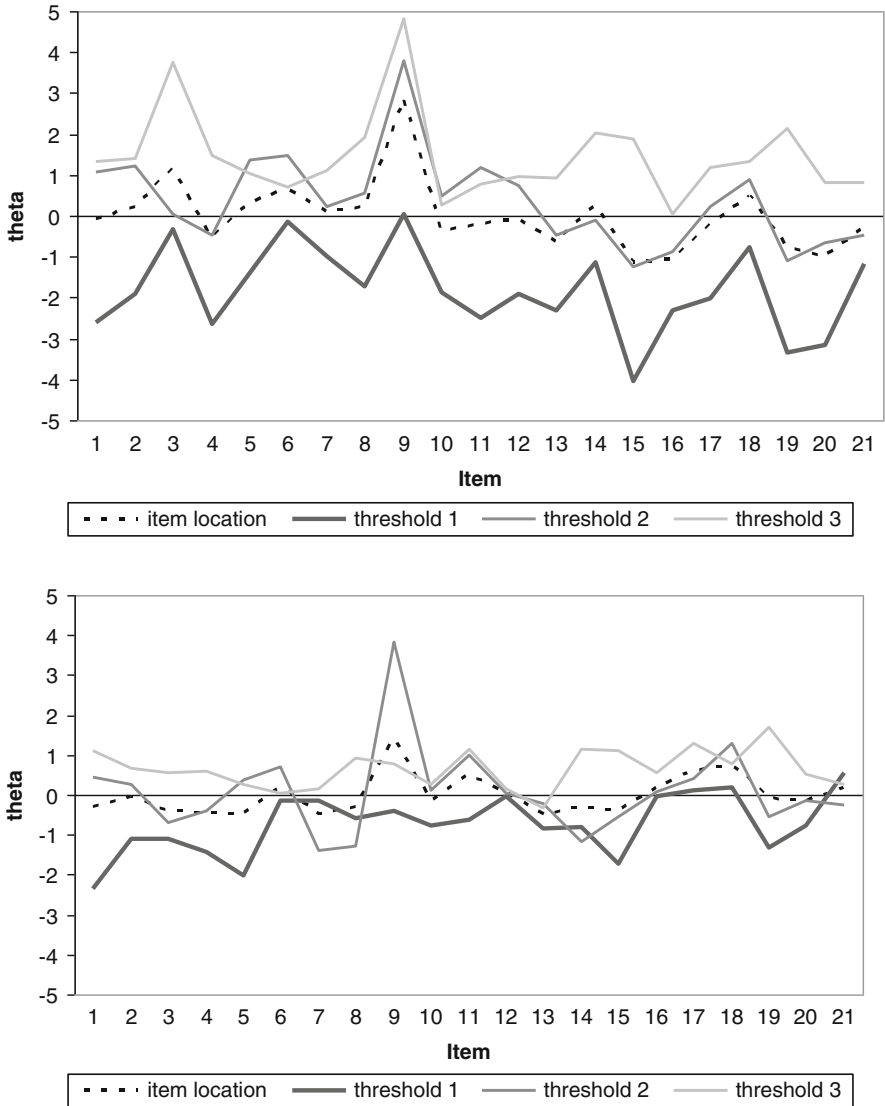


Fig. 2 Threshold parameters and item locations for the unconstrained PCM with two latent classes for the admission data (*upper part*: class 1 (MRS), *lower part*: class 2 (ERS))

for the category 3 are low. The average class membership probabilities indicate good separation in assignment of the individuals to the classes (.935 for class 1 and .907 for class 2).

For the discharge data, the BIC still favours a two-class solution (Modelfit_{2Class}: empirical $p = .032$), while the CAIC suggests a solution with only one class (Modelfit_{1Class}: empirical $p = .008$). Since the BIC is usually used as a decision

criterion and also the model-fit statistics favour the 2-RM solution, we selected the PCM with two classes. The class sizes are quite similar (63.2 and 36.7 %) compared to the admission data and also the patterns of the thresholds, indicating a class with tendency to the middle categories (class 1) and a class with tendency to the extreme categories (class 2), although the range of the thresholds increased. Concerning average class membership probabilities, the values are even better than those of the admission data (.944 for class 1 and .924 for class 2).

The constrained PCMs reveal differential results: for the admission data, the fit of the constrained PCM with two latent classes is worse than the unconstrained 2-RM, indicating additional heterogeneity; for the discharge data, the constrained and the unconstrained PCM with two latent classes are similar in model fit, especially in the BIC, indicating no additional heterogeneity.

The mean BDI-II scores at t_1 are different for the two classes ($t = -5.00$, $df = 623.5$, $p < .001$; Cohen's $d = 0.32$), with class 1 (MRS) having a mean score of 20.2 ($SD = 9.0$) and class 2 (ERS) having 23.7 ($SD = 12.8$). At discharge, the difference is larger ($t = -23.7$, $df = 526.6$, $p < .001$; Cohen's $d = 1.57$). Class 1 (MRS) has a low mean value of 5.3 ($SD = 4.2$), whereas the ERS class has a mean value of 16.0 ($SD = 8.7$).

Mixed PCM Estimated for the Virtual Sample

To assess possible qualitative change of response styles across the two time points we applied the MRM on the long format of data. The lower part of Table 1 contains also the indices for the PCM when applied to the virtual sample (long-format, left panel of Fig. 1). Both information criteria, BIC and CAIC, favor a three-class solution (Modelfit_{3Class}: $p = 0.03$). Inspection of the threshold parameters indicates that the largest class has many unordered thresholds; this class has also a mean raw score of 6.7 ($SD = 5.7$). The other two classes can be interpreted as before: class 2 seems to have a tendency to the middle categories (MRS), and class 3 prefers the extreme values (ERS). The mean class membership probability is sufficient to good with .939, .905, and .892, respectively.

Stability of Class Membership in the Virtual Sample

The members of class 1 show high stability, most of them (93.9 %) stay in the class 1 (see, Table 2). This class, however, is characterized by many unordered thresholds, and inspection of the mean BDI-II scores for this class revealed a low mean value (6.7) suggesting that the higher categories of the BDI-II items are rarely endorsed. Separating the mean BDI-II values for admission and discharge, this class has a mean sum score of 8.7 ($SD = 6.1$) at admission and of 4.8 ($SD = 3.5$) at discharge; that is, this class contains patients with low depression values at admission and even lower ones at discharge.

Table 2 Cross-classification from t_1 to t_2 in the long-format MRM with three latent classes

Class assignment at admission (t_1)	Class assignment at discharge (t_2)			Total
	Class 1	Class 2	Class 3	
Class 1	124 (93.9 %)	7 (5.3 %)	1 (0.8 %)	132 (100 %)
Class 2	510 (72.8 %)	148 (21.1 %)	43 (6.1 %)	701 (100 %)
Class 3	207 (62.5 %)	43 (13.0 %)	81 (24.5 %)	331 (100 %)
Total	841 (72.3 %)	198 (17.0 %)	125 (10.7 %)	1164 (100 %)

Note. Class assignments are given as frequencies and percentages

The majority of patients who are in the response style classes 2 or 3 at admission also move to class 1 at discharge (72.8 or 62.5 %). Obviously, class 1 consists of the much improved patients, but improvement is also remarkable in the other two classes: class 2 has a mean sum score of 21.0 ($SD = 8.6$) at admission and of 8.4 ($SD = 7.3$) at discharge; the values for class 3 are 27.3 ($SD = 11.1$) at admission and 12.4 ($SD = 9.6$) at discharge. Aside from that trend into the low symptom class 1, there is a clear preference to stay in class 2 or in class 3 and not to switch to the respective other response style class. The odds ratio for these four cells (“22,” “23,” “32,” “33”) is 6.48 (95 %-CI: 3.92–10.7).

Associations Between Latent Classes and Gender and Age

The cross-classification of gender and the assigned three classes for the long-format gives no significant association, neither at t_1 ($\chi^2 = 0.88$, $df = 2$, n.s.) nor at t_2 ($\chi^2 = 1.93$, $df = 2$, n.s.). The same is true for the separate analysis of t_1 ($\chi^2 = 2.53$, $df = 1$, n.s.); there is an association for t_2 ($\chi^2 = 5.24$, $df = 1$, $p = .022$) with female patients being underrepresented in class 1 (MRS; 62.3 % vs. 69.0 % in class 2 (ERS)), but effect size is low ($\Phi = .067$).

Concerning age, there are significant mean differences between the three classes assigned by the long-format analysis ($F(2,1161) = 14.0$, $p < .001$; $\eta^2 = .024$; mean values are 43.9, 46.6 and 43.0 years for the three classes). For the separate analysis of t_1 , there is a significant difference in age as well ($t = 5.44$, $df = 1162$, $p < .001$; Cohen’s $d = .34$). Class 1 (MRS) is slightly older with a mean value of 46.5 years ($SD = 10.4$) than the ERS class which has a mean value of 42.9 years ($SD = 11.0$).

Associations to Diagnostic Subgroups

The proportion of MRS and ERS at admission is not evenly distributed across diagnostic subgroups (see Table 3). There is preponderance for ERS in individuals with personality disorders, eating disorders, PTSD, and substance-related disorders. Patients with depression are the only group which are overrepresented in the MRS class. The remaining diagnostic subgroups (anxiety, somatoform disorders) are about uniformly distributed.

Table 3 Distribution of response style classes for diagnostic subgroups at admission

Diagnosis (ICD chapter)	Total frequency and percentage	Percentage in class 1 (MRS) and class 2 (ERS)		Odds ratio	95 % CI	
					Lower	Upper
F1 (substance-related disorders)	<i>N</i> = 254 (21.8 %)	18.9 %	27.3 %	1.62	1.22	2.15
F3 (depression)	<i>N</i> = 1001 (86.0 %)	87.9 %	82.5 %	0.65	0.47	0.91
F4 (anxiety)	<i>N</i> = 226 (19.4 %)	18.9 %	20.4 %	1.11	0.82	1.50
F4 (PTSD)	<i>N</i> = 218 (18.7 %)	15.7 %	24.4 %	1.73	1.28	2.34
F4 (somatoform disorders)	<i>N</i> = 177 (15.2 %)	16.0 %	13.8 %	0.84	0.60	1.19
F5 (eating disorders)	<i>N</i> = 212 (18.2 %)	15.0 %	24.1 %	1.80	1.33	2.43
F6 (personality disorders)	<i>N</i> = 211 (18.1 %)	12.7 %	28.3 %	2.73	2.01	3.69

Discussion

The current study examined the existence and the stability of the MRS and the ERS response styles with an IRT based approach. For this purpose the mixed PCM was used which combines the Rasch model with latent class analysis. Usually this model is used for the assessment of latent classes in which the Rasch model holds for the data. There are also studies in which the model is used for the assessment of different response styles. There are also applications testing the consistency across several traits and in longitudinal studies, but not for the assessment of response styles across time. Furthermore, our study is more complex than a simple longitudinal study, since we examined response styles in the clinical context in which mentally ill individuals received clinical intervention between the measurement points. For this purpose we used the BDI-II, a questionnaire to assess the severity of depression. For the decision on the number of latent classes, a bootstrap analysis of model fit showed always low fit values and was not very helpful; thus, this decision was based on information criteria.

The application of the mixed PCM shows interesting results for the BDI-II. The main results can be summarized as follows: For the separate analysis of the admission data (t_1), a distinction into two latent classes could be found. The classes could be interpreted as MRS and ERS. Thus, the response styles ERS and MRS that have repeatedly been found in personality and achievement tests could also be replicated with a self-report questionnaire in depression research. The constrained model fitted worse than the unconstrained model; that is, there might be some additional heterogeneity between classes beyond the response style alone (although the differences in mean BDI-II sum score are small).

For the discharge data (t_2), the separation into two latent classes indicating MRS and ERS was questionable. Furthermore, the response style classes seem to be highly confounded with depression severity when comparing the mean sum scores of the two classes. The comparison of the fit of the constrained and the

unconstrained mixed PCM with two classes, however, shows minimal differences; that is, homogeneity can be assumed. In sum, it might be concluded that the model with two classes is probably not necessary and the PCM holds for the discharge data, supporting the finding of Keller (2012) where the PCM showed the best fit in the sample of healthy individuals.

The analysis with the long-format data yields three classes, where one class contains the patients with low depression values and the other two can, again, be described as MRS and ERS. The low symptom class 1 is the largest class at discharge because most of them stay within this class and the major part of the patients in the initial classes 2 and 3 move to the class 1. Within the classes 2 and 3, there is a pronounced stability to stay, i.e., to remain in the same response style. Although additional heterogeneity has to be assumed (the constrained PCM fits worse than the unconstrained PCM with three latent classes), we may take this as a confirmation of the stability of the ERS and MRS response styles over time, as has been found before by Weijters et al. (2010) with a quite different methodological approach (the authors used a second order factor model in which they specified time-invariant and time-specific response style factors based on a coding scheme for weighting the item categories).

There are no significant relations between response style classes and gender except for the separate analysis at discharge, but effect size is low and we may conclude that gender is not related to response style to a relevant degree. However, the small effect would be in line with Weijters et al. (2010) who found that female respondents showed significantly higher levels of ERS. In contrast, Khorramdel and von Davier (2014) found no significant gender differences with regard to ERS and MRS, but their sample of students was relatively homogeneous in age and education.

The difference in age between response style classes was significant, but small in effect size and seems therefore also to be negligible. The uneven distribution in several diagnostic subgroups is an interesting result, but due to the lack of previous findings in the literature, interpretations derived only from clinical impressions may be currently too speculative before replication of these differences.

The emergence of response styles at admission and in the combined sample (long format) has implications for clinical treatment as well as for the evaluation of treatment. For treatment assignment based on the admission BDI-II score, consider a patient with a sum score of 20 which is a commonly used inclusion criterion for depression treatment studies (and may be used also in assigning treatment modules in a psychiatric/psychosomatic clinic). The corresponding person parameter in the PCM would be -0.78 ; with the additional knowledge of the response style of an individual as provided by the mixed Rasch model, the individual in the ERS class would receive a person parameter of -0.97 , while the individual assigned to the MRS class would receive a value of -0.69 . For a sum score of 14 (= cutoff for mild depression), the difference would be even larger: -1.69 for the ERS class and -1.13 for the MRS class.

In extension to this cross-sectional differential assignment of patients, one is usually interested in whether a patient has significantly improved during the stay in a clinic/from a treatment approach. One of the most popular approaches is the

Reliable Change Index (RCI—Jacobson & Truax 1991) that is based on classical test theory. Brouwer, Meijer, and Zevalkink (2013b) compare the RCI with an IRT-based change index. For a majority of cases the IRT-based statistic resulted in a similar conclusion as compared to the use of the RCI, but for some patients within the range of lower or higher change scores, IRT provided a more accurate tool (Brouwer et al. 2013b). The addition of response style information may further improve the classification into improved vs. unchanged patients (or deteriorated patients).

Currently, however, our MRM results are explorative and need to be replicated in other samples. Furthermore, other IRT-related methodological possibilities for the assessment of response styles could be examined. Multi-process IRT models have been developed and applied to decompose observed rating data into multiple response processes (Khorramdel & von Davier 2014; Plieninger & Meiser 2014). Wetzel et al. (2013) suggest conceiving response styles as their own dimension in a multidimensional model (e.g., the multidimensional random coefficient multinomial logit model by Adams, Wilson, & Wang 1997). For the purpose of measuring change, e.g., the evaluation of improvement of an individual during therapy, these multidimensional models seem a promising way to answer such research questions in longitudinal designs, and will be assessed in further studies.

Acknowledgement We thank Dr. Robert Mestel, head of Research/Quality Assurance of HELIOS Klinik Bad Grönenbach, for providing us with the dataset.

References

- Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficient multinomial logit model. *Applied Psychological Measurement*, 21(1), 1–23.
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723.
- American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.
- Baghaei, P., & Carstensen, C. H. (2013). Fitting the mixed Rasch model to a reading comprehension test: Identifying reader types. *Practical Assessment, Research & Evaluation*, 18(5). Retrieved from <http://pareonline.net/getvn.asp?v=18&n=5>.
- Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Beck depression inventory—Second edition. Manual*. San Antonio, TX: The Psychological Corporation.
- Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytic extensions. *Psychometrika*, 52(3), 345–370.
- Brouwer, D., Meijer, R. R., & Zevalkink, J. (2013a). On the factor structure of the Beck Depression Inventory-II: G is the key. *Psychological Assessment*, 25(1), 136–145.
- Brouwer, D., Meijer, R. R., & Zevalkink, J. (2013b). Measuring individual significant change on the Beck Depression Inventory-II through IRT-based statistics. *Psychotherapy Research*, 23(5), 489–501.
- Bühler, J., Keller, F., & Läge, D. (2014). Activation as an overlooked factor in the BDI-II: A factor model based on core symptoms and qualitative aspects of depression. *Psychological Assessment*, 26(3), 970–979.
- Eid, M., & Zickar, M. J. (2010). Detecting response styles and faking in personality and organizational assessments by mixed Rasch models. In M. von Davier & C. H. Carstensen

- (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 255–270). New York, NY: Springer.
- Embretson, S. E. (2010). Mixed Rasch models for measurement in cognitive psychology. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 235–254). New York, NY: Springer.
- Glück, J., & Spiel, C. (1997). Item response models for repeated measures designs: Application and limitation of four different approaches. *Methods of Psychological Research*, 2(1). Retrieved from <http://www.dgps.de/fachgruppen/methoden/mpr-online/issue2/art6/article.html>.
- Glück, J., & Spiel, C. (2010). Studying development via item response models: A wide range of potential uses. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 281–292). New York, NY: Springer.
- Gollwitzer, M., Eid, M., & Jürgensen, R. (2005). Response styles in the assessment of anger expression. *Psychological Assessment*, 17(1), 56–69.
- Hautzinger, M., Keller, F., & Kühner, C. (2006). *BDI-II. Beck depressions inventar revision—Manual [BDI-II. Revision of the Beck Depression Inventory—Manual]*. Frankfurt, Germany: Harcourt Test Services.
- Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, 59(1), 12–19.
- Keller, F. (2012). Das Beck-Depressions-Inventar (BDI-II): Psychometrische Analysen mit probabilistischen Testmodellen [The Beck-Depression-Inventory (BDI-II): Psychometric analyses with probabilistic test models]. In W. Baros & J. Rost (Eds.), *Natur- und kulturwissenschaftliche Perspektiven in der Psychologie [Natural science and cultural studies perspectives in psychology]* (pp. 120–132). Berlin, Germany: Verlag irena regener.
- Keller, F., & Kempf, W. (1997). Some latent trait and latent class analyses of the Beck-Depression-Inventory (BDI). In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 314–323). Münster, Germany: Waxmann.
- Khorramdel, L., & von Davier, M. (2014). Measuring response styles across the big five: A multiscale extension of an approach using multinomial processing trees. *Multivariate Behavioral Research*, 49(2), 161–177.
- Langeheine, R., van de Pol, F., & Pannekoek, J. (1996). Bootstrapping goodness-of-fit-measures in categorical data analysis. *Sociological Methods & Research*, 24(4), 492–516.
- Meiser, T. (2010). Rasch models for longitudinal data. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 191–200). New York, NY: Springer.
- Meiser, T., Hein-Eggers, M., Rompe, P., & Rudinger, G. (1995). Analyzing homogeneity and heterogeneity of change using Rasch and latent class models: A comparative and integrative approach. *Applied Psychological Measurement*, 19(4), 377–391.
- Meiser, T., Stern, E., & Langeheine, R. (1998). Latent change in discrete data: Unidimensional, multidimensional, and mixture distribution Rasch models for the analysis of repeated observations. *Methods of Psychological Research Online*, 3(2), 75–93. Retrieved <http://www.dgps.de/fachgruppen/methoden/mpr-online/issue5/art6/meiser.pdf>.
- Plieninger, H., & Meiser, T. (2014). Validity of multiprocess IRT models for separating content and response styles. *Educational and Psychological Measurement*. doi:10.1177/0013164413514998.
- Preinerstorfer, D., & Formann, A. K. (2012). Parameter recovery and model selection in mixed Rasch models. *British Journal of Mathematical and Statistical Psychology*, 65(2), 251–262.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.
- Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, 14(3), 271–282.
- Rost, J. (1991). A logistic mixture distribution model for polytomous item responses. *The British Journal for Mathematical and Statistical Psychology*, 44(1), 75–92.

- Rost, J. (2004). *Lehrbuch Testtheorie—Testkonstruktion [Testtheory—Testconstruction]*. Bern, Germany: Verlag Hans Huber.
- Rost, J., Carstensen, C. H., & von Davier, M. (1997). Applying the mixed Rasch model to personality questionnaires. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 324–332). Münster, Germany: Waxmann.
- Rost, J., Carstensen, C. H., & von Davier, M. (1999). Sind die Big Five Rasch-skalierbar? Eine Reanalyse der NEO-FFI-Normierungsdaten [Are the Big Five Rasch scalable? A reanalysis of the NEO-FFI norm data]. *Diagnostica*, 45(3), 119–127.
- Rost, J., & von Davier, M. (1995). Mixture distribution Rasch models. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 257–268). New York, NY: Springer.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2), 461–462.
- Spiel, C., & Glück, J. (1998). Item response models for assessing change in dichotomous items. *International Journal of Behavioral Development*, 22(3), 517–536.
- Spiel, C., & Glück, J. (2008). A model-based test of competence profile and competence level in deductive reasoning. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 45–65). Cambridge, MA: Hogrefe.
- Von Davier, M. (2001). WINMIRA 2001 user's guide. Kiel: IPN.
- Von Davier, M., & Rost, J. (1995). Polytomous mixed Rasch models. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 371–379). New York, NY: Springer.
- Ward, L. C. (2006). Comparison of factor structure models for the Beck Depression Inventory-II. *Psychological Assessment*, 18(1), 81–88.
- Weijters, B., Geuens, M., & Schillewaert, N. (2010). The stability of individual response styles. *Psychological Methods*, 15(1), 96–110.
- Wetzel, E., Carstensen, C. H., & Böhnke, J. R. (2013). Consistency of extreme response style and non-extreme response style across traits. *Journal of Research in Personality*, 47(2), 178–189.