TPM®

THE MASLACH BURNOUT INVENTORY: A TEST DIMENSIONALITY ASSESSMENT VIA ITEM RESPONSE THEORY

PASCAL JORDAN
INSTITUTE FOR OCCUPATIONAL AND MARITIME MEDICINE (ZFAM),
UNIVERSITY MEDICAL CENTER HAMBURG-EPPENDORF
UNIVERSIY OF HAMBURG

ULRICH STEINGEN
FRESENIUS UNIVERSITY OF APPLIED SCIENCES, HAMBURG

CLAUDIA TERSCHÜREN
VOLKER HARTH
INSTITUTE FOR OCCUPATIONAL AND MARITIME MEDICINE (ZFAM),
UNIVERSITY MEDICAL CENTER HAMBURG-EPPENDORF

The Maslach Burnout Inventory (MBI; Maslach, Jackson, & Leiter, 1996) is the most commonly applied measurement instrument for the assessment of burnout. In order to gain further insight into the controversy surrounding its psychometric properties, we conducted a comprehensive, robust, item response theory based psychometric analysis of a version of the German translation of the MBI within a sample of physicians recruited via a sampling procedure from a full medical register. The psychometric analysis is based on a two-step procedure which incorporates a general nonparametric analysis followed by a parametric analysis. The analysis shows departures of the MBI scales from the usually reported three-dimensional factor structure and highlights two potentially misfitting items. Deviations of the subscales from a graded response model are assessed via the use of posterior predictive *p*-values and results of a simple nonparametric method for scale construction are reported.

Key words: Burnout; Physicians; Item response theory (IRT); Mokken scale analysis; Bayesian statistics; Representative sample.

Correspondence concerning this article should be addressed to Pascal Jordan, Faculty of Psychology and Human Movement Science, University of Hamburg, Von-Melle Park 5, 20146 Hamburg, Germany. Email: pascal.jordan@uni-hamburg.de

In contrast to clinical phenomena such as depression for which an adequate, normed method of assessment, the Diagnostic and Statistical Manual of Mental Disorders criteria, exists, the assessment of burnout still lacks such a unified approach. This could be related to the severe difficulties which arise with respect to differential diagnosis (see Korczak, Kister, & Huber, 2010). As a consequence of this and due to the lack of a gold standard of measuring burnout, the construct of burnout has traditionally been equated (at least within the research domain) with the scores of a particular questionnaire, the Maslach Burnout Inventory (we refer to the Maslach Burnout Inventory-Human Services Survey, MBI-HSS — hereafter abbreviated to MBI;

TPM Vol. 25, No. 1, March 2018 – 101-120 – doi:10.4473/TPM25.1.6 – © 2018 Cises

Maslach, Jackson, & Leiter, 1996). The MBI is by far the most employed scale for the assessment of burnout (Schaufeli & van Dierendonck, 2000). Using field research along with exploratory analysis, it was constructed according to intuitive reasoning and the method of inductive item selection (Maslach, Jackson, & Leiter, 1997). More specifically, the original item pool consisting of 47 items was reduced to 22 items and each item was assigned to one of three subscales. For the latter assignment the method of (orthogonal) factor analysis was used (Maslach, Jackson, & Leiter, 1997). In a purely operational sense (for a general methodological treatment of the operational view, see Borsboom, 2006) these three subscale scores may serve the purpose of *defining* burnout. However, as the authors used a factor analysis model and as the major paradigm in evaluating psychological tests and questionnaires is based on latent variable models (see Molenaar, 1995), the MBI subscales have to be judged in terms of their psychometric properties. A comprehensive overview of conducted analyses of the scale is beyond the scope of this paper. However, it is safe to say that although some validity results (in terms of correlations with other criteria which are deemed appropriate) have been presented (e.g., Maslach, Jackson, & Leiter, 1997; Schaufeli & Enzmann, 1998), it is still — despite its popularity — an open question as to whether the scale exhibits reasonable psychometric properties and may be used in a multitude of (diagnostic) settings. More specifically:

a) The discriminatory power of the scale has been called into question. Using MBI scores as a diagnostic tool can lead to unsatisfactory sensitivity and specificity values (see, Kleijweg, Verbraak, & van Dijk, 2013). Further, Korczak et al. (2010) summarize the current status of constructing reasonable cut-offs and conclude that there is a general lack of (scientific) justification for using these cut-offs.

b) In some cases, a reduction of the scale has been suggested. In one extreme case, the reduction to a bare minimum of two items was equally successful for predictive purposes when compared to the usage of the full 22-item scale (West, Dyrbye, Satele, Sloan, & Shanafelt, 2012). In addition, single item "versions" with acceptable sensitivity (relative to the MBI) have been proposed (Dolan et al., 2015).

c) In many cases reliability estimates were barely satisfactory, passing the 0.70 threshold (Wheeler, Vassar, Worley, & Barnes, 2011). However, the generally recommended 0.80 threshold (Nunnally & Bernstein, 1995) should be passed in order to use the measurement instrument for diagnostic purposes. Note that an estimated reliability near 0.70 implies that the critical difference between two different individuals tested with the MBI is near $\sigma \cdot 1.96 \cdot \sqrt{2(1-0.7)} = 1.52 \cdot \sigma$, that is, that only a scale score difference as large as 1.5 times the standard deviation of the scale score is statistically meaningful. For a more thorough discussion of reliability thresholds we refer the reader to Carretero-Dios and Pérez (2007).

d) There has been differing evidence when it comes to the factorial structure of the MBI. Some studies showed evidence for a three-dimension structure (e.g., Hwang, Scherer, & Ainina, 2003), some reported item misfit within the three-dimension structure (Byrne, 1993), and others although supportive of a three-dimension structure — note that if indeed a two-factor solution were appropriate, then a confirmatory factor analysis (CFA) conducted under a three-factor model would not show any misfit — suggest a high correlation between two factors, so that collapsibility to a two-factor model would seem possible (e.g., Lee & Ashforth, 1990; Loera, Converso, & Viotti, 2014).

In this paper we will concentrate our efforts on d) although we will also consider and discuss some aspects of b) and c).

Note that — apart from some rare exceptions (e.g., Brand-Labuschagne, Mostert, Rothmann Jr., & Rothmann, 2013; de Vos et al., 2016; Gustavsson, Hallsten, & Rudman, 2010; Periard, 2016) — the primary approach in establishing the structure of the scale, in estimating the reliability, and in checking for model fit has been based on the classical test theory (CTT) model and the closely related approach of factor analysis. We were not able to identify a single study addressing the MBI which applies the more appropriate method of item response theory (IRT) and which at the same time relies on a valid sampling scheme. By the latter we refer to an actual sampling of subjects from a well-defined population in contrast to the commonly applied usage of ad hoc samples (e.g., graduate students). The aim of the present study is to provide an IRT analysis of the MBI questionnaire within a large sample of physicians (N = 994; Richter, Kostova, Baur, & Wegner, 2014) who were drawn from a full medical register.

After a methodological motivation of our approach in the subsequent section, we outline the employed IRT approach. The latter includes 1) a nonparametric procedure to check for test dimensionality; 2) the analysis of IRT based discrimination; 3) the highlighting of misfitted items; and 4) the application of a parametric IRT model in order to establish a common scale for further purposes (such as, e.g., linkage analysis). In the final section we summarize our findings concerning the psychometric properties with primary emphasis on test dimensionality — i.e., on point d). We provide links to CTT-based approaches and research findings and also point out methodological pitfalls concerning the usage of the MBI to measure burnout in a longitudinal framework.

METHODS

Statistical Analysis

Before describing our procedure of analysis, it is necessary to explain the advantages of an item response analysis as compared to the ordinary approach which rests on a combination of CTT and factor analysis (connections between the two approaches are highlighted in Holland & Hoskens, 2003).

The General Theoretical Framework

Drawbacks of the classical approach

One major reason for adopting IRT models instead of CTT approaches is tied to the measurement level of the items which are used in self-assessment questionnaires. The ordinal measurement scale of the items does not fit into the modeling assumptions underlying ordinary (linear) factor analysis and CTT (see Chapter 6 in Lee, 2007; Olsson, 1979b). This aspect is, for example, reflected in the partial inappropriateness of item-total correlations or in erroneous inferences regarding test dimensionality (Olsson, 1979a, 1979b). A major problem when applying CTT model to categorical response data is that quite often the items' difficulty becomes inter-

mingled with the items' discrimination. That is, in the CTT model the separation of difficulty and discrimination is presupposed, whereas within a categorical item the items' difficulty constraints the maximal achievable correlation with other items (see Appendix for an example). One well-known consequence of this, during the stage of item selection and scale construction, is the elimination of easy and difficult items simply due to the fact that their corresponding correlation with the total score is low — demonstrating the partial inappropriateness of the item-total correlation for item selection within a pool of items which vary in item difficulty. The general advice of using CTT models only within a set of items wherein respondents primarily choose the intermediate categories rests on this fact (and in addition on the underlying normality assumption). However, a scale intended to measure a latent construct should also be able to reflect differences in low or high latent ability domains and therefore needs to incorporate items which vary in difficulty.

But not only has the application of the CTT model to response data been criticized, moreover the interpretation and usage of coefficients derived from the CTT framework — like Cronbach's alpha — has been heavily called into question (Sijtsma, 2009). In diagnostic settings the use of this and other CTT-based coefficients leads to confidence intervals with uniform length — regardless of the position of the test taker on the latent continuum. In contrast IRT-based approaches rest on test information curves (Birnbaum, 1968). They therefore reflect the fact that the discriminative power of a scale is not uniform across the latent continuum.

In order to circumvent these drawbacks and to lay out a more suited method of analyzing scales, we subsequently adopt the IRT framework. Further, within this IRT-framework we try to make the analysis as robust as possible by using a set of minimal assumptions.

The approach from item response theory

Unidimensional IRT assumes that there is a latent variable θ, which is usually viewed as representing a meaningful construct (in the context of the first MBI subscale: emotional exhaustion) that determines the responses of the test taker in a probabilistic manner. If X_i denotes the response of a randomly chosen test taker to the i-th item of a scale consisting of k items, then the probability of achieving a certain score j on this item is modeled as a function of θ. The most general class of unidimensional ordinal item response models is characterized by the following modeling assumptions (Sijtsma & Molenaar, 2002):

U) There is a one-dimensional (latent) variable θ determining the responses to the items in a probabilistic manner. (Unidimensional construct)

M) The item step response functions — that is, the functions of type $P(X_i \geq j|\theta) =: f_{i,\geq j}(\theta)$ — are nondecreasing in θ. (Monotonicity)

LI) Responses to different items are stochastically independent given θ, that is, if ($i_1 < \cdots < i_l$) is an arbitrary ordered selection of test items and if $j_1, \ldots j_l$ are (any) fixed item scores, then the equation $P(X_{i1} = j_1, \ldots X_{il} = j_l|\theta) = \Pi^l_{v=1} P(X_{iv} = j_v|\theta)$ holds. (Local independence)

From these three assumptions a multitude of empirical restrictions on the data arise: for example, a) item pair covariances are guaranteed to be nonnegative; b) conditional item pair covariances are guaranteed to be nonnegative; and c) pairwise item covariances are nonnegative after stratification by the rest score (= sum score of the scale after removal of the item pair). For a more exhaustive overview we refer the reader to Holland and Rosenbaum (1986), Rosenbaum (1984), and Sijtsma and Meijer (2007).

These empirical consequences can be checked via direct computation of the (conditional) covariances for the dataset at hand without having to fit a computationally instable model or without having to rely on strong additional assumptions (like linearity). In addition, there are coefficients within the current modeling framework which allow for the quantification of item discrimination while dealing with the restriction issue on the item pair correlations resulting from the categorical measurement scale.

More specifically, the (pairwise) scalability coefficient $H_{i,j}$ (Loevinger, 1948) — playing a key role in Mokken (1971) scale analysis — is defined as the ratio of the correlation between item *i* and item *j* and the maximal conceivable correlation between two items with the same marginal distributions. The scalability coefficient H_i of item *i* is the IRT analogue of the CTT item-total correlation which however does not share with the latter the problem of mixing up item difficulty with item discrimination. It is a weighted average of the pairwise scalability coefficients involving item *i*:

$$H_i = \sum_{j \neq i} c_j H_{ij} \, , c_j := \frac{\mathrm{Cov}_{\max}\left(X_i, X_j\right)}{\sum_{l \neq i} \mathrm{Cov}_{\max}\left(X_i, X_l\right)} \tag{1}$$

Moreover, the H_i-coefficients can be combined to a total coefficient *H* which reflects the power of the whole scale. A general rule of thumb demands $H \geq 0.3$ (Mokken, 1971) in order for the scale to be considered as useful.[1]

Thus, apart from procedures which allow for the testing of the unidimensionality axioms, there are also tools for item selection available. The merit of this approach lies in its simplicity and in the absence of strong modeling assumptions which may interfere with the goal of checking for unidimensionality. That is, if, for example, the fit of a graded response model were used to examine the unidimensionality hypothesis, then a bad fit of the model would not be conclusive evidence against unidimensionality because the misfit of the graded response model could simply be caused by the deviation of the item-step-response functions from the shape of a probit curve (the usual factor analysis model shares this drawback in that it presupposes a particular shape of the conditional expected value of the item score).

After unidimensional scales have been established (or more precisely: the hypothesis of unidimensionality has not been rejected), the researcher may go on to test for specific parametric structures. This may be necessary for a multitude of reasons:

a) A comparison of scale scores across different populations of test takers is more straightforward if a common scale for the item and the person parameter has been established.

b) Linking scores across different measurement instruments (e.g., comparing burnout measurements resulting from two different scales) requires a parametric model structure.

c) More elaborate measures of item discrimination and reliability are available via the fitting of parametric models. In particular, the item and test information curves (Birnbaum, 1968) are available and enable an adaptive usage of the test as well as a quantification of the uncertainty underlying a diagnosis.

Although the fitting of (ordinary) parametric IRT model is straightforward (relying, e.g., on a very general applicable Markov chain Monte Carlo (MCMC) approach), the evaluation of the fit of an IRT model has always been a difficult task, because — if one excludes the most benevolent case of a Rasch model — there has been a lack of test statistics suited to check for mod-

el fit (a notable exception appeared only recently; see Haberman, Sinharay, & Chon, 2013). This lack can be attributed to the difficulties in establishing asymptotic results in the IRT framework. In this paper we adopt the approach of Sinharay (2005) and evaluate the fit of the IRT model without the need to refer to asymptotic statistics by applying a Bayesian approach. The latter can be characterized by the following steps:

1. For the dataset at hand, estimate the model parameters of the IRT model (e.g., the graded response model).
2. Generate $N = 1000$ pseudoreplicates of the dataset by using the estimated model parameters (more precisely by using draws from the posterior distribution of model parameters) in Step 1 and a random number generator.
3. Compare the value of an appropriately chosen test statistic — for example, item-total-correlation — for the dataset with the corresponding values of the replicates (see Subsection "Evaluating the Fit of a Graded Response Model" within Section "Results" for a description of the test statistics which were used for the MBI).
4. If the value of the present dataset is in the tail of the distribution of the replicates values, then reject the hypothesis of item/model fit.

Application to the MBI

Following the above general outline of the theoretical framework, we now describe the procedure which was used in this paper to analyze the MBI. First of all, we applied Mokken scale analysis to group the MBI items into homogeneous subscales using the previously described nonparametric measures of item discrimination H_i. The resulting subscales were then contrasted with the proclaimed three-dimension structure of the MBI and the scalability coefficients, which are not distorted by variations in item difficulty according to the remarks of the previous section, were used to further discern the items. Moreover, the property of manifest monotonicity (Junker & Sijtsma, 2000) was checked via the usage of the (default) implementation in the R package Mokken (van der Ark, 2007). Manifest monotonicity refers to the monotonicity of the conditional distribution function stated under M) — except that the conditioning is on the rest score rather than on the unobservable latent variable.

Secondly, the proclaimed subscales of the MBI were checked for unidimensionality using a nonparametric evaluation — a), b), and c) in Subsection "The approach from item response theory" — based on the computation of conditional associations (conditional association represents the key statistical property resulting from a unidimensional model; see Holland & Rosenbaum, 1986). More specifically, for each item triplet, X, Y, Z, the conditional covariances $\mathrm{Cov}(X, Y|Z = k)$ were checked descriptively for violations of unidimensionality (bearing in mind problems arising from sampling error). This provides a robust method to check for unidimensionality — as this check has to hold irrespective of the distribution of the latent variable and irrespective of the particular underlying IRT model. That is, if the model satisfies U), M), and LI) as stated in Subsection "The approach from item response theory," then the items are conditionally associated. Consequently, conditional covariance needs to be nonnegative, providing the motivation for our approach.

Finally, the last step in our approach involved the fitting of a specific parametric model to the three subscales of the MBI which were established in previous research via classical test theo-

ry methods. More precisely, we evaluated the fit of a graded response model. The reason for choosing the graded response model (apart from its popularity) is that it enables comparisons to models which use factor analysis on the polychoric correlation matrix. As already described in Subsection "The approach from item response theory," we used a Bayesian approach for fitting and evaluating the model.

For each subscale a graded response model was fitted (detailed output available on request from the corresponding author). To evaluate the model fit, we used a method which relies on drawing Monte Carlo samples (Sinharay, 2005) and which was already briefly introduced in the previous subsection. Estimated parameters of a graded response model (for the particular subscale) were used to generate new datasets (of the same size as the original one) which follow a graded response model. After N new datasets had been sampled, a test statistic was computed for each dataset and compared to the statistic corresponding to the actual dataset. If the latter was in the tail of the distribution of the simulated test statistics, then a potential misfit of the model was highlighted. For technical details and an exhaustive description of the procedure we refer the reader to Sinharay (2005).

Motivated by our aim to check for departure from unidimensionality, we used the number of conditional covariance sign violations for each item as a test statistic.[2] That is, for each sampled dataset and for each item we calculated the number of times the item was involved in a negative covariance with another item (conditionally on the score of a third item). The posterior predictive p-value corresponding to this item fit statistic is then given by $2 \cdot \min(P(T \leq t_{obs}), P(T \geq t_{obs}))$, wherein $P(T \leq t_{obs})$ is the value of the empirical distribution function of the drawn samples at the observed value of the test statistic t_{obs} for the actual dataset. Additionally, we used the quantiles of the sum score distribution as a second quantity to check for unusual deviations of the sum score distribution from the distribution implied by a graded response model.

All simulations used five chains, 6,000 iterations, a burn-in of 2,000, and a total amount of retained samples of 1,000 (convergence was assessed via graphical methods; summaries are available on request from the corresponding author). Further, a truncated normal prior with mean equal to one and unit standard deviation was used for the discrimination parameters and the prior for the threshold parameters was equal to the distribution of the order statistic derived from sampling independent standard normal variates. All computations were done in R (R Core Team, 2014) and WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000).

Participants and Instruments

In order to assess the effects of the European Working Time Directive on German hospital physicians a study assessing various variables concerning family background, workability (WAI), burnout symptoms, and working conditions (e.g., night and shift work) was conducted. The participants for the study were sampled from the full medical register of physicians working in Hamburg. That is, the population was clearly defined (all physicians working in Hamburg with entries in the register) and a sample was drawn from this population (details can be found in Richter et al., 2014). Each of the $N = 994$ participants had to complete a comprehensive questionnaire. For our current purpose we focus on the measurement of burnout and the psychometric properties of the scale, so we refer the reader to Richter et al. for a detailed description of other key variables of the study. For the measurement of burnout, the study used a version of the German translation of the Maslach Burnout

Inventory (Barth, 1985; Maslach & Jackson, 1981) consisting of the three subscales — emotional exhaustion (EE), depersonalization (DP), and lack of professional efficacy (PA) — containing nine, eight, and five items, respectively. The whole scale therefore comprises 22 items (with each item assigned to one and only one subscale) with an ordinal response format, wherein each item score ranges from 0 (*never*) to 6 (*every day*). For ease of interpretation, we reversed the coding of the PA subscale, so that high values in any subscale indicate "unfavorable" outcomes.

RESULTS

A brief overview of some sample characteristics is given in Table 1. A more comprehensive description of the dataset can be found in Richter et al. (2014). Here we focus solely on the measurement properties of the items of the MBI. Table 2 shows item-total correlations along with the reliability estimates of the subscales (we only included coefficient α because it has become common practice to report it). The reliability estimates are satisfactory — however the assumptions underlying the use of coefficient α have not been examined yet. Even more importantly, the mere fact that the reliability is high/acceptable does not tell us anything about the dimensional structure of the scale (Sijtsma, 2009). This is a common misinterpretation of α. Alpha provides a lower bound to the reliability of a test when the assumptions of CTT are met. It does not provide further clues about test dimensionality.

TABLE 1
Descriptive statistics for the variables age, gender, and position

Statistic	Age (value)	Gender	Frequency (%)	Position	Frequency (%)
Minimum	26.0	Male	55.3	Assistant doctor	67.4
1st quantile	33.0	Female	44.7	Assistant medical director	26.9
Median	39.0			Chief physician	3.9
Mean	40.5			Other	1.7
3rd quantile	47.0			Missing	0.1
Maximum	64.0				
Missing	14.0				

In fact, we think that an appropriate judgement of scale reliability should rely on the concept of test information which is founded in IRT. So, we do not dwell on the interpretation of this CTT-based summary statistic.

Finally, the histograms of the scale sum scores depicted in Figure 1 should give the reader an impression about the variation of the scale score within the sample. Note that the skewness of the scores is incompatible with the classical maximum likelihood estimation (MLE)-based factor estimation which presumes normality (see Chapter 9 in Mardia, Kent, & Bibby, 1979). Moreover, the multivariate skewness measures of 10.7 and 10.6 for the EE and the PA subscales, respectively, represent a borderline case with respect to the robustness of the MLE approach (see Muthén & Kaplan, 1985), while the corresponding skewness of 3.5 for the DP subscale is of minor concern.

TABLE 2
Correlations of each item with the rest score (= computation of the sum score after removal of the item). Reliability estimates based on coefficient α are given in brackets.

Emotional exhaustion (α = .88)	Correlation
1 Emotionally drained	.77
2 Feel used up	.61
3 Fatigued in the morning	.69
6 Stress working with people	.47
8 Burned out	.79
13 Feel frustrated	.67
14 Working too hard	.61
16 Working with patients is a strain	.43
20 Feel like at the end of the rope	.64
Depersonalization (α = .75)	**Correlation**
5 Impersonal objects	.58
10 Become more callous	.68
11 Worry about hardening emotionally	.63
15 Don't care about patients	.50
22 Patients blame for their problems	.24
Personal accomplishment (α = .81):	**Correlation**
4 Understand patients	.47
7 Deal with problems	.56
9 Positive influence	.62
12 Energetic	.47
17 Create relaxed atmosphere	.57
18 Exhilarated after work	.53
19 Accomplishment	.54
21 Deal with emotional problems calmly	.42

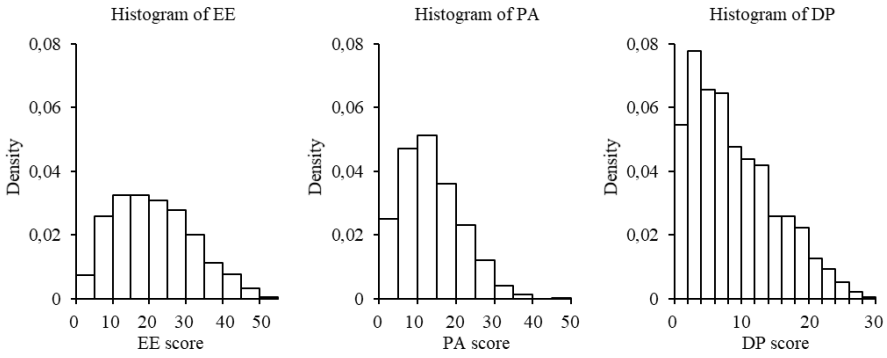

FIGURE 1
Histogram of the subscale scores of the MBI (German version).
EE = emotional exhaustion; PA = personal accomplishment; DP = depersonalization.

Mokken Scale Analysis

Following the concepts outlined in Section "Methods," we first report the scalability coefficients — we used the R-package Mokken (Van der Ark, 2007) for computation — H_i for the items of the MBI under the given assignment to the three subscales. As can be seen in Table 3, the scalability coefficients within the PA subscale are of similar magnitude, whereas the items of the EE and of the DP subscales show stronger variations in scalability. In particular Item 22 is marked as a low discriminating item within the DP subscale. Overall the relative ordering of the items within each subscale is comparable to the ordering given by the item-total-correlation in Table 2. The total scalability coefficients are 0.509, 0.367, 0.397 for the EE, the PA, and the DP, respectively. In addition to the scalability coefficients, Mokken scale analysis includes a check of nonnegativity of the pairwise correlations and a check of monotonicity. The pairwise correlations are all nonnegative. The check of manifest monotonicity (necessary for unidimensionality) shows some violations for the PA and the DP subscales. These violations (see column labeled "#vi" in Table 3) are primarily located within the items with lowest H_i. Especially the last two items show strong violations of monotonicity.

TABLE 3
Nonparametric scalability coefficients H_i (along with their estimated asymptotic standard errors) for the items of the MBI subscales. The number of violations of the monotonicity of the function $P(X_i \geq j|R_{-i} = r)$ is shown in the column labeled #vi

EE	H_i (*SE*)	#vi	PA	H_i (*SE*)	#vi	DP	H_i (*SE*)	#vi
Item 1	0.590 (0.015)	0	Item 4	0.347 (0.024)	0	Item 5	0.432 (0.020)	0
Item 2	0.519 (0.019)	1	Item 7	0.390 (0.021)	0	Item 10	0.486 (0.018)	0
Item 3	0.531 (0.017)	0	Item 9	0.424 (0.018)	0	Item 11	0.462 (0.021)	0
Item 6	0.383 (0.024)	0	Item 12	0.335 (0.022)	4	Item 15	0.385 (0.024)	1
Item 8	0.604 (0.014)	0	Item 17	0.392 (0.022)	0	Item 22	0.199 (0.028)	3
Item 13	0.517 (0.018)	0	Item 18	0.371 (0.022)	1			
Item 14	0.488 (0.019)	0	Item 19	0.374 (0.021)	2			
Item 16	0.381 (0.028)	1	Item 21	0.300 (0.023)	6			
Item 20	0.525 (0.020)	0						

Note. H_i = scalability coefficient; *SE* = standard error; #vi = number of violations of monotonicity; EE = emotional exhaustion; PA = personal accomplishment; DP = depersonalization.

The values of the scalability coefficients in Table 3 are dependent on the actual assignment of the items to the subscales. Consequently, they do not indicate if an item is assigned to the "most suited" subscale. We now contrast the results of a nonparametric item selection procedure based on Mokken scale analysis (see Straat, Van der Ark, & Sijtsma, 2013) with the hypothesized assignment of the items. The procedure starts with the item pair with maximal pairwise scalability coefficient and then seeks to add new items (ensuring maximal scalability coefficients) to the scale until the total *H* index drops below .30. As can be gleaned from Table 4 only two scales are established and two of the 22 items are deemed as unscalable. Moreover, literally every item of the DP subscale is collapsed into the EE subscale — that is, according to the item selection procedure, the EE and the DP subscale are indistinguishable.

We note however, that the newly created Scale 1 (see Table 4) still shows violations of manifest monotonicity and that increasing the cut-off in the automatic selection procedure did not resolve this issue (extensive summary statistics of the manifest monotonicity checks are available on request from the corresponding author).

TABLE 4
Scalability coefficients (H_i) and number of violations of manifest monotonicity (#vi) after applying an automated nonparametric item selection procedure

Scale 1	H_i (*SE*)	#vi	Scale 2	H_i (*SE*)	#vi
Item 1	0.503 (0.016)	0	Item 4	0.387 (0.026)	0
Item 2	0.437 (0.019)	0	Item 7	0.414 (0.024)	0
Item 3	0.459 (0.016)	1	Item 9	0.459 (0.020)	1
Item 5	0.336 (0.020)	0	Item 17	0.399 (0.024)	0
Item 6	0.345 (0.021)	1	Item 18	0.413 (0.023)	0
Item 8	0.519 (0.015)	0	Item 19	0.411 (0.023)	1
Item 10	0.390 (0.020)	4			
Item 11	0.439 (0.020)	1			
Item 12	0.354 (0.020)	2			
Item 13	0.474 (0.016)	1			
Item 14	0.406 (0.018)	0			
Item 15	0.324 (0.021)	5			
Item 16	0.373 (0.023)	1			
Item 20	0.444 (0.019)	0			

Note. H_i = scalability coefficient; *SE* = standard error; #vi = number of violations of monotonicity.

Nonparametric Assessment of the Unidimensionality of the Subscales

Until now, we have only contrasted the actual assignment of items with the results of a nonparametric item selection procedure. We now head toward a more rigorous approach for the testing of test dimensionality. More specifically, we first report the results of a nonparametric procedure, as described in Subsection "Application to the MBI," to check the unidimensionality of each scale and then proceed with the fitting of parametric item response models. Figure 2 depicts the relative frequencies of conditional covariance sign violations for each subscale of the MBI. As can be seen from Figure 2, the nonparametric check for unidimensionality casts doubt on the unidimensionality of the EE and of the DP subscales. As already emphasized in Subsection "Application to the MBI," if a unidimensional IRT model holds, then all pairwise conditional item covariances need to be nonnegative — this follows from Lemma 2 in Lehmann (1966) and an inequality of Shea (1979); see also Holland & Rosenbaum (1986) and Rosenbaum (1984). Now, given the large amount of computable covariances, some are expected to be negative by chance (due to sampling error). However, a proportion of above 15% for Items 2, 16, and 20 for the EE as well as for the last item of the DP seems very questionable. In contrast, the results for the PA are of no severe concern (all frequencies lie below 10%; Items 4 and 21 exhibit the highest amount of violations: 7-8%). In conjunction with the

scales resulting from the automated item selection procedure, it can however also be noted that this check for unidimensionality reveals violations. For example, Items 2, 5, 12, and 20 of Scale 1 each show relative frequencies of approximately 15%. On the other hand, Scale 2 seems satisfactory with all relative frequencies below 2%.

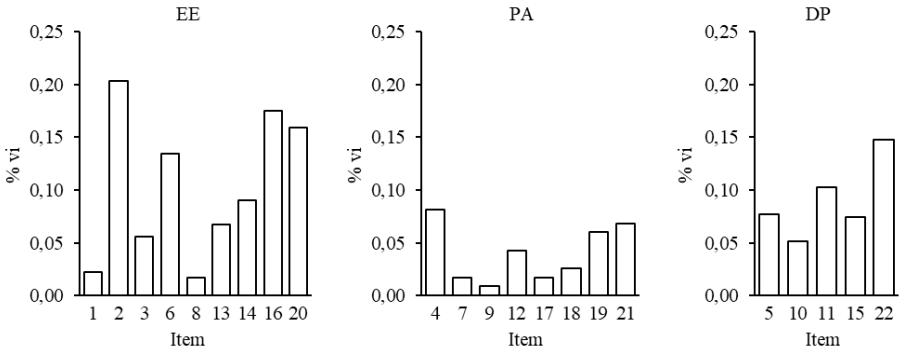

FIGURE 2
Percentages of conditional covariance sign violations for the MBI subscales.
EE = emotional exhaustion; PA = personal accomplishment; DP = depersonalization.

Evaluating the Fit of a Graded Response Model

We now look at the model fit of a graded response model to each of the subscales. The results of the testing procedure which was outlined toward the end of Subsection "Application to the MBI" are reported in Table 5. Note that Table 5A refers to item-specific diagnostics. It reports the percentage of datasets (sampled with parameter draws from the posterior) which show more extreme conditional covariance sign violations (involving the particular item) than the observed dataset. Table 5B does not report item-specific checks but aims at diagnosing violations of the normality assumption of the latent variable. For each quantile listed in the left column, the percentages of posterior simulated datasets (they, by definition, are all generated using a normal distribution for the latent variable) that result in more extreme values for the quantile of the total sum score, are reported. As can be seen from Table 5A and Table 5B, the item fit for the items of the EE subscale is unsatisfactory, in that Item 2 and Item 20 show posterior predictive *p*-values of zero, indicating that not even a single sampled dataset showed more extreme values on the respective item specific test statistic than the observed data.

For the PA and DP subscales at least one item exhibits a low posterior predictive *p*-value. With respect to the examination of the distribution of the sum score, each subscale shows strong deviations when compared to data sampled from a graded response model. The latter may indicate either a misfit of the item step response function or a deviation of the distribution of the latent variable from normality.

We also evaluated the fit of a graded response model for the two subscales of the automated item selection procedure. Scale 1 contained five items with a PPV of zero and Scale 2 contained two items with a PPV of zero. Therefore, neither of the two subscales shows satisfactory model fit.[3]

TABLE 5A
Posterior predictive *p*-values (PPV) for the number of conditional covariance sign violations

EE		PA		DP	
Item	PPV	Item	PPV	Item	PPV
Item 1	0.43	Item 4	0.08	Item 5	0.42
Item 2	0.00	Item 7	0.48	Item 10	0.73
Item 3	0.16	Item 9	0.51	Item 11	0.37
Item 6	0.43	Item 12	0.48	Item 15	0.23
Item 8	0.55	Item 17	0.51	Item 22	0.08
Item 13	0.11	Item 18	0.30		
Item 14	0.10	Item 19	0.01		
Item 16	0.51	Item 21	0.27		
Item 20	0.00				

Note. PPV = posterior predictive *p*-values; EE = emotional exhaustion; PA = personal accomplishment; DP = depersonalization.

TABLE 5B
Posterior predictive *p*-values (PPV) for the quantiles of the sum score

Quantiles	EE	PA	DP
	PPV	PPV	PPV
0.1	0.01	0.13	0.00
0.2	0.04	0.00	0.02
0.3	0.03	0.00	0.00
0.4	0.04	0.00	0.02
0.5	0.07	0.00	0.00
0.6	0.06	0.00	0.00
0.7	0.02	0.01	0.00
0.8	0.03	0.02	0.02
0.9	0.04	0.01	0.30

Note. PPV = posterior predictive *p*-values; EE = emotional exhaustion; PA = personal accomplishment; DP = depersonalization.

Confirmatory Factor Analysis

For the sake of completeness, we also report the results of CFA for a three-factor model with the postulated item structure (to ensure identifiability one can either fix the variances of the latent factors or the loadings of the first items of the respective subscales) and unrestricted covariances between the factors. The confidence interval for the root mean square error of approximation (RMSEA) value was [0.078, 0.086] indicating misfit of the model. In addition, we have standardized root mean square residual (SRMR) = .087 and comparative fit index (CFI) = .842 (checking the unidimensionality of each subscale separately did also not result in a satisfactory model fit).

DISCUSSION

We analyzed the structure of the MBI scale within a well-defined population of physicians, a population wherein the phenomenon of burnout is quite common and highly relevant. Based on the results of the IRT analysis we may conclude that the postulated factor structure of the MBI is questionable — at least within the underlying population our sample refers to. First of all, the fit of a graded response model was rejected for all subscales (as well as the fit of the ordinary linear factor analysis model). Although this does not necessarily imply a deviation from unidimensionality of the subscales (as the rejection could just be based on item step response functions deviating from probit/logistic/linear shape), some evidence via the use of flexible nonparametric approaches has been provided to even question a common metric for the items of each subscale. One further key advantage of this nonparametric check is that it enables an evaluation of the measurement instruments (i.e., the items) which is not distorted by any (potentially wrong) assumptions of normality of the latent variable (Holland & Rosenbaum, 1986; Rosenbaum, 1984). As can be seen by inspection of Table 5B, a violation of the normality assumption is not implausible — thus, it is even more important to scrutinize the results from the view of nonparametric IRT. In fact, if a scale passes the nonparametric check then — although specific parametric models may fail to fit — ordering of the subjects with respect to the simple sum score is justified (a detailed account of this ordering property can be found in Hemker, Sijtsma, Molenaar, & Junker, 1997). Based on the results depicted in Figure 2, the general unidimensionality hypothesis seems questionable as there are significant proportions of negative conditional covariance pairings (the PA subscale showing perhaps the least severe violations). Note that this sort of violation is quite a severe violation as it implies the following: stratify the population according to the performance on an item (say: “Working too hard”). Then choose one stratum (the units within each stratum show the same response on the item “Working too hard”) and calculate within this stratum the correlation between the responses on two additional items (say: “Feel used up” and “Feel frustrated”). Then the indicated violation, that is a negative conditional covariance, implies that within the stratum those physicians who are feeling more frustrated are feeling less used up. Of course, overall, that is unstratified, there is a positive correlation between the two items. However, conditioning on the performance on one item, eliminates/reverses this relationship. The latter phenomenon is also known under the label “Simpsons paradox” (see Agresti, 2002) and must not occur if the items of the scale are supposed to measure a common construct (all correlations — whether conditional or unconditional — must show the same sign; see Holland & Rosenbaum, 1986). Moreover, if one takes into consideration the additional aspect of discriminative power (a concept which is not directly related to the dimensionality issue) then at least two items (Items 21 and 22) show quite unsatisfactory values — casting doubt on their usefulness — even if they were fitting within a unidimensional framework.

The application of a nonparametric item selection procedure (Straat et al., 2013) did result in just two scales (with basically the EE and DP subscales collapsed into a single scale). This is in accordance with some previous results which noted a high correlation between two factors within the three-dimension CFA model or which directly showed that a third factor is superfluous (de Vos et al., 2016). However, the results of this procedure should be interpreted with caution as there is no rigorous control of strict unidimensionality within this automated item selection procedure (the mere fact that each pairwise scalability coefficient is nonnegative is necessary but far

from being sufficient for the hypothesis of unidimensionality to hold (see Rosenbaum, 1984). In fact, only Scale 2 looks promising as Scale 1 did not withstand the conditional covariance check. Overall, we may conclude that there is ample and robust evidence — from four different approaches (conditional covariance approach, Mokken scale analysis, graded response model testing, and to some extend even the CFA framework) — that the presumed scale structure does not hold.

Indeed, in many studies which employed the traditional CFA approach, an acceptable model fit of the three-factor structure could only be achieved by either a) allowing for correlated measurement errors (i.e., violating the fundamental principle of local independence); b) eliminating items; or c) allowing for substantial deviations from simple structure. Moreover, there seems to be a lack of consistency (see Loera et al., 2014) in that, for example, the items which were eliminated in one study, were retained in another study. Of course, this points to substantial sample and population dependency of the properties of the scale. In fact, to the authors' knowledge there has not been a single study which tested and confirmed measurement equivalence of the MBI between different populations in a strict sense. Most studies addressed the consistency of the number of underlying factors, a much weaker criterion of comparability than measurement equivalence.

As already mentioned in the introduction, only a rather limited number of studies used the IRT paradigm instead of the CFA/CTT framework. Thus, comparisons of our results are somewhat limited in this regard. In our opinion, the best reference regarding the implications of IRT is the dissertation of Periard (2016) because it i) provides an in-depth analysis of the MBI using parametric IRT models and the concept of item and test information curves; ii) it relies on a very large sample size. The main conclusion, which is consistent with our findings, is that the three subscales of the MBI should not be used for scoring and that the traditional three-factor model is inappropriate.

Before concluding the discussion referring to test dimensionality it is important to point out our focus on testing unidimensionality of each subscale rather than fitting a common three-dimension joint model. This approach is justifiable by the following key observation: if a three-dimension model with the proposed simple structure (i.e., each item is assigned to exactly one dimension) holds, then three separate unidimensional models need also hold. That is, the unidimensionality of each subscale is necessary for the existence of the presumed three-dimension model. If violations at the level of the subscales can be pointed out, then a fortiori the proclaimed higher dimensional model does not hold. The check of the sign of conditional covariances provides the most important observation with respect to this, as it rests on minimal assumptions (i.e., any undimensional model, regardless of the precise model specification, needs to pass this check).

The conducted analyses also attempted to look at modified versions of the subscales in order to arrive at unidimensional scales. However, as already noted, the first scale of the automated item selection procedure does not pass the required checks — in fact, manifest monotonicity, conditional covariances, and the graded response posterior predictive check all indicate violations of similar size as already noted for the EE subscale. Scale 2 does much better in this respect: only the graded response model fit fails. Now, a comparison of Table 4 with the original setup of the MBI subscales shows that Scale 2 is basically a reduced PA subscale. The scale is reduced by exactly those two items exhibiting the most severe violations of manifest monotonicity as depicted in Table 3. Thus, it might be stated that a reduced PA subscale passes nonparamet-

ric unidimensionality checks and therefore allows for the ordering of subjects with respect to their sum score. We note that an analogous attempt to improve the fit of the DP subscale by elimination of the last item does not work (e.g., no substantial reduction in relative frequencies of conditional covariance sign).

As the question of predictive validity is concerned, we note — in conjunction with the results of West et al. (2012) — that one item of the whole EE subscale accounts for $0.79^2 = 62{,}4\%$ of the total rest score variance (see Table 2). It appears that just asking whether or not a person feels burned-out accounts for more than half of the combined variation provided by the sum score of the remaining items of the EE subscale. In fact, the explained variance exceeds 82% by just using the first and the fifth item (i.e., Item 1 and Item 8). When combined with the hypothesis that item scalability decreases toward the end of the questionnaire, one could make a valid point for a reduction of the scale.

The preceding remarks only addressed the cross-sectional aspect of the measurement of burnout, where psychometric limitations of the scale were pointed out. The question concerning the longitudinal measurement of burnout was not addressed (and cannot be addressed with the current dataset). However, as in many diagnostic settings the scale will ultimately be used to highlight time-dependent development of burnout — we finally want to emphasize an important additional pitfall which needs to be avoided when employing the MBI for these purposes. As Molenaar (2004) clearly demonstrated, the usual psychometric analysis of a scale only addresses the level of interindividual comparisons and — apart from some rare exceptions (ergodic processes) — does not have any bearing on the psychometric structure in terms of intraindividual comparisons (closely related results are discussed in Steingen, 2013). That is, even if the factor structure of the MBI were three dimensional (on an interindividual level), there is no guarantee that the same factors (and even the same number of factors) could be identified when considering longitudinal data (i.e., data wherein individual trajectories are provided). This practically important step of extending the analysis to the intraindividual level is seldom considered and should, in our opinion, be one focus of further research.

NOTES

1. If a scale adheres to the unidimensionality hypothesis then it can be shown that all scalability coefficients are nonnegative. The rule reflects this property but also adds the requirement that $H \geq 0.3$, so that unidimensional scales with a large amount of "noise" are also deemed as inappropriate — although technically speaking the unidimensionality hypothesis would hold).
2. Zhu and Stone (2011) show via simulation that in general, statistics which are focused on pairwise association are more appropriate for checking departures from unidimensionality than item-level statistics such as item-total correlations.
3. Note that although a method based on Bayesian posterior predictive *p*-values is conceptually very different from the classical frequentistic approach, results of simulation studies (Zhu & Stone, 2011) indicate that an application of this Bayesian method does not result in increased type 1 error rates — in fact, we also ran a small simulation of $n = 100$ datasets generated according to a graded response model (with test length equal to nine and varying parameter values) and noticed a total of 54 rejections out of 900 ($= 9 \times 100$) conducted tests, when using the conventional $\alpha = 5\%$ threshold.

REFERENCES

Agresti, A. (2002). *Categorical data analysis* (2nd. ed.). New Jersey, NJ: John Wiley & Sons.

Barth, A. (1985). *Das MBI-D: Erste Untersuchung mit einer deutschen Übersetzung des Maslach Burnout Inventory bei klientenzentrierten Gesprächstherapeuten und Hochschullehrern* [The MBI-D: First analysis of client-centered talking therapists and professors using a German translation of the Maslach Burnout Inventory]. Unpublished manuscript, Zulassungsarbeit zur staatlichen Ergänzungsprüfung im Fach Psychologie mit Schulpsychologischem Schwerpunkt, University of Erlangen-Nürnberg, Germany.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 379-397). Reading, MA: Addison-Wesley.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*, 425-440. doi:10.1007/s11336-006-1447-6

Brand-Labuschagne, L., Mostert, K., Rothmann, S. R., Jr., & Rothmann, J. C. (2013). Burnout and work engagement of South African blue-collar workers: The development of a new scale. *Southern African Business Review*, *16*(1), 58-93.

Byrne, B. M. (1993). The Maslach Burnout Inventory: Testing for factorial validity and invariance across elementary, intermediate and secondary teachers. *Journal of Occupational and Organizational Psychology*, *66*, 197-212. doi:10.1111/j.2044-8325.1993.tb00532.x

Carretero-Dios, H., & Pérez, C. (2007). Standards for the development and review of instrumental studies: Considerations about test selection in psychological research. *International Journal of Clinical and Health Psychology*, *7*(3), 863-882.

de Vos, J. A., Brouwers, A., Schoot, T., Pat-El, R., Verboon, P., & Näring, G. (2016). Early career burnout among Dutch nurses: A process captured in a Rasch model. *Burnout Research*, *3*, 55-62. doi:10.1016/j.burn.2016.06.001

Dolan, E. D., Mohr, D., Lempa, M., Joos, S., Fihn, S. D., Nelson, K. M., & Helfrich, C. D. (2015). Using a single item to measure burnout in primary care staff: A psychometric evaluation. *Journal of General Internal Medicine*, *30*, 582-587. doi:10.1007/s11606-014-3112-6

Gustavsson, J. P., Hallsten, L., & Rudman, A. (2010). Early career burnout among nurses: Modelling a hypothesized process using an item response approach. *International Journal of Nursing Studies*, *47*, 864-875. doi:10.1016/j.ijnurstu.2009.12.007

Haberman, S. J., Sinharay, S., & Chon, K. H. (2013). Assessing item fit for unidimensional item response theory models using residuals from estimated item response functions. *Psychometrika*, *78*, 417-440. doi:10.1007/s11336-012-9305-1

Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika*, 62, 331-347. doi:10.1007/BF02294555

Holland, P. W., & Hoskens, M. (2003). Classical test theory as a first-order item response theory: Application to true-score prediction from a possibly nonparallel test. *Psychometrika*, *68*, 123-149. doi:10.1007/BF02296657

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics*, *14*(4), 1523-1543.

Hwang, C. E., Scherer, R. F., & Ainina, M. (2003). Utilizing the Maslach Burnout Inventory in cross-cultural research. *International Journal of Management*, *20*(1), 3-10.

Junker, B. W., & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement*, *24*, 65-81. doi:10.1177/01466216000241004

Kleijweg, J. H., Verbraak, M. J., & van Dijk, M. K. (2013). The clinical utility of the Maslach Burnout Inventory in a clinical population. *Psychological Assessment*, *25*, 435-441. doi:10.1037/a0031334

Korczak, D., Kister, C., & Huber, B. (2010). *Differentialdiagnostik des Burnout-Syndroms* [Differential diagnosis of the burnout syndrome]. *Schriftenreihe Health Technology Assessment (HTA) in der Bundesrepublik Deutschland* (Bd 105). Köln, Germany: Deutsches Institut für Medizinische Dokumentation und Information (DIMDI).

Lee, R. T., & Ashforth, B. E. (1990). On the meaning of Maslach's three dimensions of burnout. *Journal of Applied Psychology*, *75*(6), 743-747.

Lee, S. Y. (2007). *Structural equation modeling: A Bayesian approach*. New York, NY: Wiley.

Lehmann, E. L. (1966). Some concepts of dependence. *Annals of Mathematical Statistics*, *37*, 1137-1153. doi:10.1214/aoms/1177699260

Loera, B., Converso, D., & Viotti, S. (2014). Evaluating the psychometric properties of the Maslach Burnout Inventory-Human Services Survey (MBI-HSS) among Italian nurses: How many factors must a researcher consider? *PLoS ONE*, *9*. doi:10.1371/journal.pone.0114987

Loevinger, J. (1948). The technic of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. *Psychological Bulletin*, *45*, 507-530. doi:10.1037/h0055827

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS-a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325-337. doi:10.1023/A:1008929526011

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate analysis*. London, UK: Academic Press.

Maslach, C., & Jackson, S. E. (1981). The measurement of experienced burnout. *Journal of Organizational Behavior*, *2*, 99-113. doi:10.1002/job.4030020205

Maslach, C., Jackson, S. E., & Leiter, M. (1996). Maslach Burnout Inventory. Manual (3rd ed.). Palo Alto, CA: Consulting Psychologists Press.

Maslach, C., Jackson, S. E., & Leiter, M. (1997). Maslach Burnout Inventory: In: C. P. Zalaquett & R. J. Wood (Eds.), *Evaluating stress: A book of resources* (pp. 191-218). Lanham, MD: Scarecrow Press.

Mokken, R. J. (1971). *A theory and procedure of scale analysis with applications in political research*. The Hague, The Netherlands: Mouton.

Molenaar, I. W. (1995). Some background for item response theory and the Rasch model. In G. Fischer & I. W. Molenaar (Eds.), *Rasch models. Foundations, recent developments, and applications* (pp. 3-14). New York, NY: Springer.

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology – This time forever. *Measurement*, *2*, 201-218. doi:10.1207/s15366359mea0204_1

Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology*, *38*, 171-189. doi:10.1111/j.2044-8317.1985.tb00832.x

Nunnally, J. C., & Bernstein, I. J. (1995). *Teoría psicométrica* [Psychometric theory]. Madrid, Spain: McGraw-Hill.

Olsson, U. (1979a). On the robustness of factor analysis against crude classification of the observations. *Multivariate Behavioral Research*, *14*, 485-500. doi:10.1207/s15327906mbr1404_7

Olsson, U. (1979b). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*, 443-460. doi:10.1007/BF02296207

Periard, D. A. (2016). *A bifactor model of burnout? An Item Response Theory analysis of the Maslach Burnout Inventory Human Services Survey* (Doctoral dissertation). Retrieved from http://corescholar.libraries.wright.edu/etd_all/1534/

R Core Team. (2014). *R: A language and environment for statistical computing*. Wien, Austria: R Foundation for Statistical Computing.

Richter, A., Kostova, P., Baur, X., & Wegner, R. (2014). Less work: More burnout? A comparison of working conditions and the risk of burnout by German physicians before and after the implementation of the EU Working Time Directive. *International Archives of Occupational and Environmental Health*, *87*, 205-215. doi:10.1007/s00420-013-0849-x

Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika*, *49*, 425-435. doi:10.1007/BF02306030

Schaufeli, W. B., & Enzmann, D. (1998). *The burnout companion to study and practice: A critical analysis*. London, UK: Taylor & Francis.

Schaufeli, W. B., & van Dierendonck, D. (2000). Utrechtse burnout schaal: Handleiding [Utrecht burnout scale: Manual]. Lisse, The Netherlands: Swets Test Publishers.

Shea, G. (1979). Monotone regression and covariance structure. *Annals of Statistics*, *7*, 1121-1126. doi:10.1214/aos/1176344794

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach alpha. *Psychometrika*, *74*, 107-120. doi:10.1007/s11336-008-9101-0

Sijtsma, K., & Meijer, R. R. (2007). Nonparametric item response theory and related topics. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, Vol. 26: Psychometrics* (pp. 719-746). Amsterdam, The Netherlands: Elsevier.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory* (Vol. 5). Thousand Oaks, CA: Sage.

Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement*, *42*, 375-394. doi:10.1111/j.1745-3984.2005.00021.x

Steingen, U. (2013). *Längsschnittforschung in der Psychologie: Eine methodenkritische Analyse am Beispiel von Burnout* [Longitudinal analysis in psychology: A methodological critical analysis via the case of burnout.] (Doctoral dissertation). Retrieved from http://ediss.sub.uni-hamburg.de/volltexte/2013/6154/ pdf/Dissertation_Steingen_Haupttext.pdf

Straat, J. H., Van der Ark, L. A., & Sijtsma, K. (2013). Comparing optimization algorithms for item selection in Mokken scale analysis. *Journal of Classification*, *30*, 72-99. doi:10.1007/s00357-013-9122-y

Van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software*, *20*, 1-19. doi:10.18637/jss.v020.i11

West, C. P., Dyrbye, L. N., Satele, D. V., Sloan, J. A., & Shanafelt, T. D. (2012). Concurrent validity of single-item measures of emotional exhaustion and depersonalization in burnout assessment. *Journal of General Internal Medicine*, *27*, 1445-1452. doi:10.1007/s11606-012-2015-7

Wheeler, D. L., Vassar, M., Worley, J. A., & Barnes, L. L. (2011). A reliability generalization meta-analysis of coefficient alpha for the Maslach Burnout Inventory. *Educational and Psychological Measurement*, *71*, 231-244. doi:10.1177/0013164410391579

Zhu, X., & Stone, C. A. (2011). Assessing fit of unidimensional graded response models using Bayesian methods. *Journal of Educational Measurement*, *48*, 81-97. doi:10.1111/j.1745-3984.2011.00132.x

APPENDIX

Range Restrictions of Correlation Coefficients Induced by Differing Marginals: An Illustration Via Binary Random Variables

In this Appendix we illustrate the restriction in the correlation coefficient which is imposed solely by differences in marginal distributions. We confine ourselves to the case of two binary random variables — but the qualitative conclusion generalizes to other settings, for example, item-total correlation of ordinal variables, as well. Thus, the purpose of this example is to demonstrate that items whose item difficulties are extreme relative to the other items of the scale will automatically be deemed as inappropriate items when judged by the ordinary item-total correlation. Suppose now, that we are given the following general cross tabulation of two binary random variables:

TABLE A1
General cross tabulation of two binary random variables

$X\backslash Y$	$Y = 1$	$Y = 0$	X-Marginal
$X = 1$	$P(X = 1, Y = 1)$	$P(X = 1, Y = 0)$	a
$X = 0$	$P(X = 0, Y = 1)$	$P(X = 0, Y = 0)$	$(1-a)$
Y-Marginal	b	$(1-b)$	1

Without loss of generality let $a < b$. Then the following entries for the cross table maximize the correlation between X and Y (while maintaining the marginal distribution):

TABLE A2
Cross tabulation with maximal correlation for the given marginal distributions in A1 (subject to $a < b$).

$X\backslash Y$	$Y = 1$	$Y = 0$	X-Marginal
$X = 1$	a	0	a
$X = 0$	$b-a$	$1-b$	$(1-a)$
Y-Marginal	b	$(1-b)$	1

That is, for any two binary random variables X, Y with given marginal probabilities $P(X = 1) = a < b = P(Y = 1)$ the maximal correlation is given by:

$$\frac{a-ab}{\sqrt{b(1-b)}\sqrt{a(1-a)}} = \frac{a(1-b)}{\sqrt{b(1-b)}\sqrt{a(1-a)}} = \frac{\sqrt{a}}{\sqrt{1-a}}\frac{\sqrt{1-b}}{\sqrt{b}}$$

As an example, let $a = 0.5$, $b = 0.9$, then insertion of these values in the above formula shows that:

$\rho_{X,Y} \leq \frac{1}{3}$