Original Article

“Use the Force!” Adaptation of Response Formats

From Rating Scale to Multidimensional Forced Choice

Alexander Leonard Schünemann and Matthias Ziegler

Institute of Psychology, Humboldt-Universität zu Berlin, Germany

Abstract: The present paper features the adaptation of an existing Big Five questionnaire with a rating scale (RS) response format into a measure using a multidimensional forced choice (MFC) response format. Rating scale response formats have been criticized for their proneness to intentional and unintentional response distortions. Multidimensional forced choice response formats were suggested as a solution to mitigate several types of response sets and response styles by design. The Big Five Inventory of Personality in Occupational Situations (B5PS) is a situation-based questionnaire designed for personnel selection and development purposes which would benefit from *fake-proof* response formats. MFC response formats require special effort during test construction and calibration which will be laid out here. Changing the response format has severe consequences on item design and scoring. An inherent issue with MFC formats derives from their inability to yield interpersonal comparative results from standard (sum) scoring. This issue can be solved with item response theory (IRT)-based calibration during test construction. The Thurstonian IRT approach (TIRT) was developed by Brown and Maydeu-Olivares (2011), and aspects of MFC item design and TIRT calibrations are explored in this paper. Evidence on structural and construct validity are presented alongside recommendations on the test development processes. The results support the feasibility of the concept of MFC test construction with TIRT calibration in a contextualized and situation-based item format.

Keywords: multidimensional forced choice, Thurstonian IRT, response format adaptation

An aspect of test construction often taken for granted without much debate is the selection of an appropriate response format. While a lot is talked about nomological networks, constructs to be measured, intended use of measures, target population, and indicators of psychometric quality, the question of which response format to use is usually answered rather intuitively by defaulting to a standard rating scale response format. Though more than one option exist, the vast majority of questionnaires feature one of any rating scale (RS) response formats. RS formats have been predominantly used in research and practice and utilized for all aspects of (personality) research even beyond psychometric assessment due to their ease of construction, scoring, and psychometric evaluation. However, one recurring topic of debate remains their susceptibility to intentional and unintentional response distortions, more commonly referred to as faking or social desirability in different settings (education, HR, clinical, etc.) and under various circumstances (e.g., low-stakes vs. high-stakes assessments). In addition, it has been shown that RS are also susceptible to response styles such as middle or extreme point responses (e.g., Wetzel et al., 2016; Ziegler & Kemper, 2013). Such phenomena have sparked a lot of research aimed at preventing or modeling these response sets and styles in order to obtain *bias*-free scores. A different approach which has yielded some success is the use of a different answer format which prevents these phenomena. Among these, multidimensional forced-choice (MFC) response formats have been suggested and tested (Brown, 2015, 2016; Brown & Maydeu-Olivares, 2011, 2012, 2013, 2018; Cao & Drasgow, 2019). The current paper presents the adaptation of an existing RS format questionnaire into an MFC version. The primal questionnaire, the Big Five Inventory of Personality in Occupational Situations (B5PS), uses RS items embedded in situational vignettes (Ziegler et al., 2019). The MFC adaptation will maintain this contextualized item design feature, setting the adapted questionnaire apart from other existing MFC questionnaires.

Introduction to the MFC Response Format

The general idea behind MFC formats as opposed to RS formats is to not let test takers respond to every test item separately on a scale from x to y but instead offer several items at once in an item block (typically three or four items are combined). Importantly, each item reflects a different dimension of the constructs the questionnaire intends to

https://doi.org/10.1027/2698-1866/a000044

measure (e.g., Big Five domains). Test takers are then *forced* to choose between the given items with a certain number of response options at their disposal, thus making it impossible to endorse all dimensions at once. Direct comparison of items removes the need for separate individual rating scales for each item. Most commonly, test-takers are asked to state which of the items within an MFC block applies most and which one applies least to them. Endorsing one option (e.g., as most like me) automatically means that none of the other items in the same block can also be rated this way. Thereby, the MFC format makes it practically impossible to present oneself in a socially desirable way on all given dimensions of the questionnaire. The question why MFC formats are not widely used can be answered easily based on two main emerging issues: (1) increased complexity in test construction and (2) ipsative data (meaning self-referencing) derived from traditional sum scoring (Meade, 2004). As opposed to normative data, which are usually generated by questionnaires utilizing RS formats (created by using norm samples based on sum scores or factor scores), ipsative data are created when sum scoring MFC questionnaires. Ipsative data are solely self-referencing in a way that test scores cannot be compared across individuals. The sum of test scores for any individual test-taker adds to a constant value (usually 0), which represents the overall distributable *points* that can be achieved on an MFC questionnaire, and which is the same exact value for everybody (Baron, 1996). Therefore, it becomes impossible to rank or compare people based on their sum scores alone since their aggregated results will be the identical, and they have no unifying common point of reference. Ipsative profiles can only be used for intrapersonal analyses, which pose a substantial issue in many fields of application for MFC questionnaires such as personnel selection where the comparability of test results is of essence. The first issue, a more complex test construction process (1), can be addressed by paying attention to several specific aspects of MFC item design, mainly the careful compilation of items within an item block, but the latter, certainly more grave issue of ipsative data (2), has proposedly been solved by several authors over the last decade by utilizing item response theory (IRT)-based modeling approaches (Brown, 2016; Brown & Maydeu-Olivares, 2011; Stark et al., 2005).

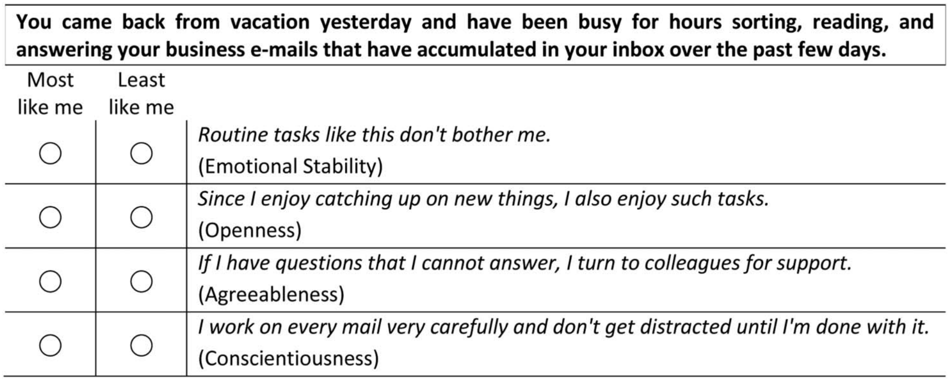
Implications for the Test Adaption

Aim of this project was to explore several aspects of MFC-specific test construction and item design issues by demonstrating an exemplary Thurstonian IRT (TIRT; Brown & Maydeu-Olivares, 2011) calibration during the adaptation of an existing Big Five inventory. TIRT calibration instead of traditional sum scoring of MFC questionnaire data is necessary to render ipsative sum scoring data back to normative test scores which then allows for interpersonal comparisons (Dueber et al., 2019). The TIRT approach by Brown and Maydeu-Olivares (2012) is based around modeling the actual response process inherent to MFC questionnaires. As opposed to RS response formats, test takers not only rate themselves in respect to one given dimension (represented by each item) at a time but engage in a comparative judgment of several items within an item block by making binary comparisons between the given stimuli (and inter-relating them). The MFC response process therefore substantially differs from RS response processes (Fuechtenhans & Brown, 2022; Sass et al. 2020). Modeling these multiple, binary comparisons to form a comparative judgment is at the core of the modeling approach of Brown and Maydeu-Olivares (2011). Detailed explanations and in-depth information on the TIRT modeling and calibration can be found in Brown (2015, 2016) and Brown and Maydeu-Olivares (2011, 2012, 2013, 2018).

The TIRT calibration has many implications for the test construction process presented in this study, and several aspects of item design were investigated during this test adaptation. Those aspects include scale or dimension count (many vs. few; Schulte et al., 2021), item block size (two to multiple; Brown & Maydeu-Olivares, 2011), item context (single statement vs. contextualized; Shaffer & Postlethwaite, 2012), permutations within blocks (Wetzel et al., 2021), the keyed direction of items (positive, negative, or mixed; Bürkner et al., 2019; Walton et al., 2020), and the question of whether full or partial ranking is used given the binary response options *most like me* and *least like me* in combination with more than three response options (Brown & Maydeu-Olivares, 2011). The aforementioned aspects impact, among other things, the overall questionnaire length and the thereby induced test-takers strain (Sass et al., 2020), trait score estimate reliability (Frick et al., 2023) and validity (e.g., convergence with RS based trait measures or external validation criteria; Walton et al., 2020; Wetzel & Frick, 2020), and ultimately the MFC questionnaires potential to reduce susceptibility to SDR (low-stakes assessments) and faking (high-stakes assessments; Cao & Drasgow, 2019). All of these aspects must be considered to successfully calibrate an MFC questionnaire to combat the main issue of ipsative versus normative data in MFC questionnaire scoring.

The Original Questionnaire

The current paper focusses on the response format transformation of the B5PS (Big Five Inventory of



You came back from vacation yesterday and have been busy for hours sorting, reading, and answering your business e-mails that have accumulated in your inbox over the past few days.		
Most like me	Least like me	
○	○	*Routine tasks like this don't bother me.* (Emotional Stability)
○	○	*Since I enjoy catching up on new things, I also enjoy such tasks.* (Openness)
○	○	*If I have questions that I cannot answer, I turn to colleagues for support.* (Agreeableness)
○	○	*I work on every mail very carefully and don't get distracted until I'm done with it.* (Conscientiousness)

Figure 1. Sample item block. Test-takers have to choose the item describing them best (most like me) and worst (least like me). In the version presented to the test takers, the constructs assessed are not listed. This information was added here to showcase the principle of combining items from different domains. The situational embedding of the item content requires test takers to at least be slightly familiar with the situational context given in each item. The majority of situation vignettes are placed in organizational settings referring mostly to white-collar working environments such as giving presentations, networking and colleague/management interactions, conferences or workshop participations, and the likes. In case no prior experience exists in relation to a certain situation, participants are asked to imagine how they would react in such situations.

Personality in Occupational Situations; Ziegler, 2014) which was specifically designed for personnel selection and development, a field of application especially prone to response distortions (social desirability vs. faking). The current endeavor has been initialized with the intention to leverage the potential of the MFC response formats while paying attention to the aspects of item construction (1) and the TIRT calibration (2). The B5PS presented itself as a suitable basis for adapting into an MFC questionnaire due to the aforementioned relevance of SDR and faking in HR contexts and also because it features a specific item design based on situational vignettes. The test taker is confronted with a work-related critical incident and then asked to rate a single statement on a 6-point rating scale in reference to the situations described before, therefore (work related) contextualizing items to increase validity (Shaffer & Postlethwaite, 2012) in its field of application. A sample item is shown in Figure 1. The B5PS features a hierarchical structure with 42 facets beneath the Big Five dimensions and has been thoroughly applied in research and practice. Findings on scale intercorrelation, reliability, structural validity, and construct validity evidence can be found in Ziegler et al. (2019). To adapt an existing rating scale questionnaire instead of starting all over from scratch for an MFC questionnaire allowed us to make use of preexisting information on psychometric indicators that guided the test adaptation process in item construction and TIRT calibration.

Proceeding from a contextualized item design of the B5PS, we intend to explore several aspects of item construction specifically relevant to MFC formats. For example, the compilation of items within an MFC block which is usually done by independently prerated SDR on single statement items (rating scale). Those are then put together in blocks of approximately equal SDR (Wetzel & Frick, 2020). In case of a contextualized embedding, all items within a block refer to a given situational vignette and are therefore much more interlinked. Designing items based on a given situational context poses a challenge to the construction of an MFC questionnaire not previously tackled. Here we propose a more thorough approach to MFC block building and pay special attention to an iterative item design process preceding the TIRT calibration work, a necessity previously pointed out before by Bürkner et al. (2019).

A major intention and challenge of this research endeavor is to keep the measured construct, the intended use, and the target population of the original measure widely intact even if the MFC response format adaptation drastically changes the item format, the response process, and the scoring algorithm. As previously mentioned, the constructs being measured are the Big Five (Goldberg, 1990) and 42 subsumed facets (Rouco et al., 2022; Ziegler et al., 2019). The intended use of the measure is allocated in personnel selection and development for organizational settings in HR contexts. The item content is based on typical work-related situations from a white-collar working environment, and therefore, the target population mostly defines itself as having experienced more years of education, an age range between 18 and 65 years, and a balanced gender distribution. This has implications for the samples used but also the validation strategy (Kemper et al. 2019; Ziegler, 2014). Specifically, the scores derived from the MFC version should yield a structure comparable to the RS version, and correlations with scores for other constructs should be comparable as well.

Aims of the Study

In this paper, we will focus on adhering to the aforementioned relevant aspects of MFC item design and their implication for TIRT calibration, striving for several goals: (1) a (acceptably) balanced distribution of responses over all items within an MFC block, achieved by iteratively adjusting item difficulties. This sets the basis for the TIRT calibration and is a first indicator of successful MFC questionnaire design; (2) to provide evidence for structural validity and therefore an intact dimension/facet hierarchy in the Big Five framework identical to the B5PS in its rating scale version. This will be approached in two ways. First, a successful TIRT calibration with the facets as latent variables will attest to the assumed allocation of items to facets. Second, a confirmatory factor analytical model using the TIRT-based factor scores will be used to underscore the assumed relations between Big Five domains and facets. Finally, (3) first evidence for convergent and discriminant validity with scores from a typical Big Five rating scale questionnaire will be provided.

Overall, this paper aims to demonstrate the feasibility to adapt, construct, and calibrate a hierarchical, situation-based Big Five inventory into an MFC format while keeping the inherent nomological network intact. Moreover, the paper reflects lessons learned on several aspects of MFC item design and the construction process and delivers a proof of concept for situation-embedded MFC response formats to be used in high-stakes assessment situations in HR contexts.

Method

Construction of the B5PS-FC

The measure constructed and used in this paper was named the B5PS-FC, being the MFC version of the B5PS(-RS) for distinction purposes. Before commencing MFC item construction, several a priori decisions had to be made in respect to the total amount of item blocks in the questionnaire, the effective block size or the number of items per item block, the compilation and permutation of items and their respective facets within a block, the homogenous or mixed keyed direction of items, and whether full or partial ranking should be employed. The decision process and its consequences on item design and TIRT calibration are elaborated in the following.

Item Design

The situation vignettes used for the situational embedding of the item blocks were taken from the B5PS. The actual items to be used in blocks were specifically designed for each situation aiming at specific domains and facets to match the situation described in the vignette. To mirror the approximate test duration and number of items (210) of the B5PS, we decided to use 50 item blocks each featuring four items reflecting four different facets. Each block was tailored to a preselected situation vignette. Thus, all four items had to be feasible responses to this specific vignette. Such a quadruple block design (4 items in every item block) also allowed for a thorough permutation of all facets within and across all item blocks. Facets were represented by four to seven items each with only “interest in reading” being an outlying low with just two representations due to the comparable lack of relevance to the overall field of application and often lack of feasibility to meaningfully integrate into the situational embedding.

Next, we decided to approach the item compilation in a different way as opposed to prior research in which items are typically combined based on prerated singular ratings of SDR usually based on rating scale scores (Wetzel et al., 2021). Given the fact that our items were supposed to be situation-embedded, all items had to refer to a given situational context and therefore could not be assembled randomly without respecting their main reference to the situation vignette at hand in every block. This becomes clearer given a sample item in Figure 1.

For comparison, the original single stimulus rating scale item from the B5PS-RS corresponding to this sample situation vignette is *I am easily distracted*. (Conscientiousness) on a six-point rating scale ranging from *strongly disagree* to *strongly agree*. The other three items added in the MFC item block are not rated in the RS version for this vignette. Figure 1 exemplifies the contextualized vignette at the top, being the point of reference for the four items below. Each response option represents one domain (and facet) of the Big Five. The test taker is then asked to select the one response option which is most like me and another one that is least like me.

Obviously, each item within the MFC item blocks has to be directly related to the situation vignette on the content level. We therefore took another route to item compilation which eludes criticism faced toward prerated SDR-based item compilation (Bürkner et al., 2019; Pavlov et al., 2021) and instead constructed specific items tailored to every given situation vignette. We iteratively adapted all items within each and every block over several trial runs during the initial test construction, challenging the item design and item block compilation repeatedly. This was then followed up by three distinct pilot phases, an initial qualitative review by (1) test construction experts, a subsequent qualitative item review by (2) HR professionals and practitioners, and eventually a quantitative analysis based on (3) a piloting sample of $n = 100$. Phase 1 encompassed experts ($n = 12$) from acquainted academic institutions willing to review our initial item pool and situation vignettes. We paid special attention to specifically require these experts to possess extensive knowledge on psychological constructs, such as the Big-Five model utilized here, and experience in test construction processes together with expertise on MFC response formats. After reworking our questionnaire based on this first qualitative feedback from an academic point of view, we invited HR professionals and practitioners to challenge our revised item pool from a real-world application perspective. Clients and customers (n = 20) who previously worked with the B5PS in organizational settings were asked to review this new MFC-based version and provide feedback on the applicability of the new MFC items for personnel selection and personnel development purposes. Participants in this pilot phase were employed in medium-to-large corporations, spread over several industries such as information and communications technology, finance, automotive, and insurance and ranged over several hierarchical levels from HR experts to head of HR positions. The aim was to gather a diverse group of professionals with a multitude of perspectives to critically revise the item design from several perspectives, including applicability, practicability, selectivity, ethical suitability, and general utility for their respective field of application. Based on the qualitative feedback collected from the first two pilot phases, several adaptations to the items were implemented. The then reworked and finalized third pilot version was subsequently administered to a sample of 100 participants to gather quantitative data on response distributions. This subsample was drawn from the same sample pool that was later used to form the larger TIRT calibration sample. This multistep piloting phase process was realized to render each item within every block as equally desirable or *difficult* as possible based on both qualitative expert ratings and quantitative data collected. The outcome of these iterative sessions of item revision and data collection were meant to even out the response pattern distributions to an acceptable level of at least a minimal number of endorsements across and a balanced most-like-me and least-like-me response count within the item blocks. It is of vital importance to a MFC questionnaire's success to offer several well-matched items, almost equal in item difficulty (probability of endorsement), in such a way that eventually only the trait standing in each dimension assessed makes the real difference between each item within a block and ultimately results in the test-takers endorsement or rejection. The distribution of response patterns is therefore a strong indicator of item quality. Assuming Big Five trait standings are normally distributed within a population, we expect to see close to even distribution across all offered items in our MFC blocks. One or two items dominating a single block while other options are fully neglected would be an indicator of item bias. The iterative process of item adjustment based on expert, practitioner, and participant data resulted in maximized item quality with a satisfactorily balanced distribution of responses over all dimensions, facets, and items within each block.

In terms of the keyed direction of items and given the highly elaborated item content due to the contextualization of item stems, we decided to only go for positively keyed items. This goes against recommendations by Brown and Maydeu-Olivares (2011) or Wetzel et al. (2021) to use mixed keyed directions of items to improve TIRT model convergence and trait recovery. However, when constructing our items, it became clear that mixed keys within a single situation vignette are extremely difficult to phrase and often entail highly unlikely behaviors or simply feel fabricated and easily identifiable. Thus, we decided to forego the recommendation for the sake of obtaining a questionnaire with more utility and acceptance in practice and to avoid overtly socially undesirable items.

In respect to whether full or partial ranking as response option, the most-like-me/least-like-me format was chosen which yields partial rankings in a quadruple block size like ours. Partial rankings were then imputed as recommended (Brown & Maydeu-Olivares, 2012).

TIRT Calibration

In essence, the TIRT calibration itself revolves around modeling the response process test takers engage in while answering MFC questionnaires. Since all items within an item block are related, test takers engage in comparative judgments of given items within an item block. This is reflected in the TIRT modeling approach by transforming response patterns into patterns of binary comparisons between constructs represented through the items within a block (for further elaboration on the basics of the TIRT scoring algorithm, see Brown, 2010). In the end (in a quadruple block size like ours with items A–D), response patterns like A > B, A > C, A > D, B > C, B > D, and C > D are formed and then scored respectively (in this case, A was selected as most-like-me and D was selected as least-like-me resulting in an answer vector 1,1,1,NA,1,1; the relation between B > C must be imputed). From a technical perspective, the modeling requires extensive coding, high computational power, and assumptions on scale correlations facilitating the convergence of the model. All this is done in Mplus (Muthén & Muthén, 1998-2019) based on an Excel macro automation provided by Brown and Maydeu-Olivares (2012) to aid with code creation. A good overview on other practical applications of TIRT modeling can be found in Wetzel et al. (2021).

To successfully calibrate a data set according to the TIRT modeling approach, several steps beyond the binarization of the response raw data must be taken which are exemplified by Brown and Maydeu-Olivares (2012). In addition, it can become necessary to impose model constraints to facilitate model convergence in case of more complex models like ours that also feature hierarchical structures (facets and domains). For example, the covariance structure between the latent variables can be specified to initially enable the estimations to converge. Following such a strategy, we a priori defined that in case of computational convergence issues, we would use theory to derive such covariance assumptions. Thus, we assumed that facets from different domains should be uncorrelated once social desirability is controlled for (Bäckström & Björklund, 2014; Ziegler & Bühner, 2009), and moreover, we assumed that facets belonging to the same domain should have covariances comparable to the ones found for the rating scale version of the questionnaire. Consequently, such estimates were used as starting values in our model to facilitate convergence, a procedure recommended by Brown and Maydeu-Olivares (2012) for more complex models like ours.

Unfortunately, the calculation of goodness-of-fit statistics for the TIRT calibration was not possible due to the sheer complexity of the modeling approach (50 item blocks with 200 items and 42 correlated latent variables) and the limited computational resources at hand to feasibly estimate *SE*s and goodness-of-fit statistics. This issue has already been reported by Brown and Maydeu-Olivares (2012), and following their recommendations, the feasibility of the model was then manually assessed by checking whether model loadings are keyed in the right direction on item level for each item within a facet in reference to all other items within their respective item block. For the selected model, factor scores were derived and used as a basis for further analyses (SEM, reliability estimates, convergent/discriminant validity) which in comparison did provide goodness-of-fit statistics unlike the TIRT calibration itself.

Data Collection

The B5PS-FC was designed as an online questionnaire and could therefore be easily administered in an unproctored web-based setting which was used for all data collections. Several measures were implemented as quality control checks during quantitative data collection which are strongly recommend for all MFC sampling procedures. Reasonable minimum response time thresholds per item block were set together with the implementations of three control questions to filter out participants who just clicked through the questionnaire without working on it. Along with that, repetitive response patterns were investigated and individual data sets were excluded in case of abnormal response behavior that would imply nonconformity with the test instruction.

During the iterative process of item development, described more elaborately in the above paragraph on item design, two qualitative feedback pilot phases were conducted in which both test development experts (n = 12) and HR professionals (n = 20) were involved and asked to provide input on item content. After adjusting the pilot questionnaire based on these two feedback loops, a third quantitative pilot phase of data collection (n = 100) was conducted to check for empirical response distributions. After no further adaptations to the questionnaire were deemed necessary, data collection continued to a total of 547 participants. Participants were gathered from an online access panel company and received monetary compensation for their effort. The target population for data collection were people in employable age ranges with higher educational levels to represent the more white-collar-oriented application of personality assessments in organizational settings. Consequently, the samples used here were drawn from the general population with limitations on age range, prior work experience, and with equal gender balance. It was sought after higher educational levels (high school degree and higher), and the amount of work experience was monitored during the main data collection. It must be noted that the MFC format inherently requires a higher cognitive understanding and self-reflection capability to be able to fulfill the task of multiple binary comparisons between items and to understand and empathize with the given situation vignette. Therefore, it was vital to draw a sample from higher educational background as a proxy for higher cognitive capability and also being a native speaker in the given language (in this case German) was a requirement, which was adhered to during data collection to accurately represent the target population.

Sample

The final sample used for the main analyses consisted of 547 participants representing the target population of the questionnaire in terms of appropriate age range, gender distribution, work experience, and educational level. Gender distribution within the sample was almost balanced with 52.3% (n = 286) of the sample being female, and age ranged from 20 to 60 with an average of 41.34 years (*SD* = 10.26). Additional inclusion/exclusion criteria were applied for educational level, requiring at least higher education entrance qualification, preferably above. 30.5%

(*n* = 166) of participants passed higher education entrance qualification while 45.3% (*n* = 247) of the sample had a university degree (Bachelor, Master, PhD). This limitation was imposed to account for the expected higher cognitive demand for *solving* MFC questionnaires as stated for example by Sass et al. (2020). Additionally, given the questionnaires setting being contextualized in organizational surroundings, prior work experience was required from all participants.

Statistical Analyses

All data preparation and analyses were conducted in R with RStudio (R Core Team, 2021, version 4.0.4; RStudio Team, 2021, version 1.4.1106), Microsoft Excel (2016), and Mplus (Muthén & Muthén, 1998–2019, Version 7.4.). The following R packages were used: *lavaan* (Rosseel, 2012), *mice* (van Buuren & Groothuis-Oudshoorn, 2011), *dplyr* (Wickham et al., 2022), *knitr* (Xie, 2022), *purrr* (Henry & Wickham, 2022), *MplusAutomation* (Hallquist & Wiley, 2018), *psych* (Revelle, 2019), and *car* (Fox & Weisberg, 2019).

Preparing and recoding of the data also included the imputation of missing by design values due to incomplete rankings in a quadruple block design with the most-like-me/least-like-me response options format. TIRT calibrations were performed in Mplus but without computation of goodness-of-fit measures or *SE* estimations due to the complexity and size of the model as recommended by Brown and Maydeu-Olivares (2012). Factor scores from the TIRT Mplus modeling were extracted to further test the factor structure (SEM) to obtain evidence for structural validity. Concerning model fit test statistics for SEM, recommendations from Hu and Bentler (1999) were adapted to take into account the high complexity of the applied models and the limited degrees of freedom due to the extensiveness of the SEM approach. Therefore, model fit was mainly judged based on the standardized root mean residual (SRMR < .11) and the comparative fit index (CFI > .95) where applicable to be more suitable for the comparably complex calculations carried out in this study. This is based on recommendations by Greiff and Heene (2017) and Kenny et al. (2015) who determined that SEM model parameter might be inflated due to either limited degrees of freedom or complexity of the utilized models. In general, the approximate model fit was gauged based on a combination of indicators and always compared to less elaborate models with slight differences in the modeling in respect to the comparative fit as suggested by Heene et al. (2011). Adjustments to the SEMs were made based on modification indices according to Saris et al. (2009) and Bentler and Chou (1992).

Research Goal A: Response Distributions

One prerequisite for a successful TIRT calibration is balanced item difficulties which result in evenly distributed response patterns. Item difficulty in respect to MFC questionnaire relates to the probability of endorsement of a given item and has to be understood in context of the entire item block presented to the test taker. The first goal of the iterative questionnaire development process was to achieve a good distribution of responses over all Big Five dimensions, aiming for an approximately 20% split of responses attributed to each Big Five dimension. Furthermore, evenly splitting response across Big Five facets was also aimed for and within each item block, it was important to balance response allocations not only between most like me and least like me on a single item but also balanced across all items within a block. This was checked based on quantitative data (*n* = 100) resulting from the third piloting phase and will be shown in the Results section Part A.

Research Goal B: Structural Validity Evidence

A major research goal is to retain the nomological net captured in the original measure while a successful response format adaptation is carried out. The successful convergence of the TIRT calibration in Mplus is an indicator that the proposed hierarchical structure of the RS version can also be found in the MFC version of the B5PS. As opposed to most Big Five MFC questionnaires, which either only model the Big Five domains or aim for a *facet only* level analyses to achieve a high scale count, we implemented a design combining both, the domain and the facet level. This required additional constraints to the TIRT modeling to realize this hierarchical structure and to facilitate model convergence. We therefore used the facet constellation and intercorrelation from the B5PS-RS version (Ziegler, 2014; Ziegler et al., 2019) to provide basis for the MFC model and implemented within-domain correlation constraints in the Mplus syntax by imposing starting values for the estimations ranging from .50 to .60, which is in line with the within-domain correlations of the B5PS-RS. This was done to facilitate model estimation for a high complexity model with large scale count in a hierarchical design, a practice recommendation by the method developers (Brown & Maydeu-Olivares, 2012) and common practice in complex models (e.g., recommended for latent change score models or moderated nonlinear factor analysis). We safely assumed that within-domain correlations can be derived from what is empirically known about the actual scale intercorrelations from the B5PS rating scale version, since we did neither intend to change the construct nor the facet to domain allocation of the B5PS during this adaptation process.

By completing the model computations in Mplus, it can be assumed that the intended hierarchical allocation of items to facets is intact in the MFC version. We therefore expect to replicate the structural validity evidence found for the scores derived from the original B5PS-RS since we do not intend to change anything about the construct itself. To this end, we will use a series of CFAs. First, we will test measurement models for each domain. In case of insufficient goodness-of-fit statistics or nonconvergence, modifications of the measurement model according to modification indices (Bentler & Chou, 1992; Saris et al., 2009) are considered within plausible ranges. Second, we will combine these domain-specific models into a joint structural model of all domains. A bifactor solution, which describes a modeling approach including a dedicated latent factor to engulf all variance attributed to social desirable responding (Bäckström & Björklund, 2014; Ziegler & Bühner, 2009), will intentionally not be used for the MFC version since the nature of the MFC format claims to mitigate or even eliminate those response sets and response styles that would normally confound test results with social desirable responding. We therefore did not specify an additional bifactor, loading all domain and facet scores, reflecting SDR which was done for the B5PS rating scale version. The MFC version model should be able to work without a bifactor and demonstrate mostly uncorrelated Big Five domains as will be analyzed as well. Additionally, analyses of measurement invariance across typical sample indicators such as age, gender, and educational level were carried out to accumulate further evidence for a successful modeling of a stable structure.

Research Goal C: Convergent and Discriminant Validity

To obtain convergent validity (and to an extent, discriminant validity as well) evidence, a rating scale Big Five measure, the Big Five Structure Inventory (BFSI/S2; Arendasy, 2011) was used. Small to moderate convergent correlations are expected, considering that the BFSI operationalizes the Big Five with a differing facet structure (Pace & Brannick, 2010). Additionally, the BFSI features a single statement and noncontextualized item format which is much more generic than the situationally embedded item design of the B5PS-FC.

Results

Response Distributions

To ensure that items within a block are balanced, meaning that they are chosen with comparable probabilities, we looked at the response distributions during the quantitative piloting phase. Tables 1–6 show that the overall response allocation for both Big Five dimensions and facets were evenly balanced. Response distributions on dimension level ranged from 18.6% to 22.7%, only deviating slightly from the expected 20% cut. Facets were expected to be chosen equally as well and proved to come in at around the expected 10%–12% range. Also, response distribution split evenly between most-like-me and least-like-me response options on almost all dimensions and facets (anything from 40% to 60% was aimed for and held in most cases). Out of the 400 response options (50 item blocks consisting of four items with two response options each), only four were not selected by any test taker. Thus, overall, the iterative balancing process yielded item blocks with satisfying selection probabilities for each item. A full comparison with the overall response distributions from the final version of the B5PS-FC based on the calibration sample can be found in the online supplementary material. Response distribution deviations from the theoretically perfect cut (20% for each Big Five domain) from the pilot phase data to the full sample ranged from Δ −1.4% (Neuroticism) to +2.7% (Conscientiousness) for the pilot data to .9% (Neuroticism) to +1.4% (Agreeableness) in the full sample which represents a satisfactory overall response distribution in both the pilot data set (*n* = 100) and the extended full sample (*N* = 547). A successful iterative item adaptation process has therefore been achieved by the multistep process utilized

Table 1. Response distribution during pilot – Big Five domain level

	Response distribution (%)			
Big Five domain	Overall	Most	Least	Δ
N | Emotional stability	928 (18.6%)	397 (42.8%)	591 (57.2%)	−134 (−14.4%)
E | Extraversion	999 (20.0%)	541 (54.2%)	458 (45.8%)	83 (8.3%)
O | Openness	916 (18.9%)	590 (64.4%)	326 (35.6%)	264 (28.8%)
A | Agreeableness	1,020 (20.4%)	443 (43.4%)	577 (56.6%)	−134 (−13.1%)
C | Conscientiousness	1,137 (22.7%)	529 (46.5%)	608 (53.5%)	−79 (−6.9%)

Table 2. Response distribution during pilot – Big Five facet level – Emotional Stability

	Response distribution (%)			
Facets	Overall	Most	Least	Δ
N | Emotional Stability				
Drive	123 (13.3%)	74 (60.2%)	49 (39.8%)	25 (20.3%)
Mental balance	112 (12.1%)	68 (60.7%)	44 (39.3%)	24 (21.4%)
Emotional robustness	119 (12.8%)	44 (37.0%)	75 (63.0%)	−31 (−26.1%)
Equanimity	130 (14.0%)	79 (60.8%)	51 (39.2%)	28 (21.5%)
Self-attention	145 (15.6%)	49 (33.8%)	96 (66.2%)	−47 (−32.4%)
Carefreeness	164 (17.7%)	26 (15.9%)	138 (84.1%)	−112 (−68.3%)
Confidence	135 (14.5%)	57 (42.2%)	78 (57.8%)	−21 (−15.6%)
E | Extraversion				
Forcefulness	94 (9.4%)	41 (43.6%)	53 (56.4%)	−12 (−12.8%)
Energy	133 (13.3%)	81 (60.9%)	52 (39.1%)	29 (21.8%)
Conviviality	112 (11.2%)	67 (59.8%)	45 (40.2%)	22 (19.6%)
Humor	76 (7.6%)	31 (40.8%)	45 (59.2%)	−14 (−18.4%)
Communicativeness	112 (11.2%)	49 (43.8%)	63 (56.3%)	−14 (−12.5%)
Sociability	113 (11.3%)	51 (45.1%)	62 (54.9%)	−11 (−9.7%)
Positive attitude	102 (10.2%)	65 (63.7%)	37 (36.3%)	28 (27.5%)
Readiness to take risks	153 (15.3%)	88 (57.5%)	65 (42.5%)	23 (15.0%)
Wish for affiliation	104 (10.4%)	68 (65.4%)	36 (34.6%)	32 (30.8%)
O| Openness				
Open-mindedness	84 (9.2%)	63 (75.0%)	21 (25.0%)	42 (50.0%)
Creativity	128 (14.0%)	100 (78.1%)	28 (21.9%)	72 (56.3%)
Intellect	115 (12.6%)	77 (67.0%)	38 (33.0%)	39 (33.9%)
Artistic interests	90 (9.8%)	56 (62.2%)	34 (37.8%)	22 (24.4%)
Willingness to learn	111 (12.1%)	60 (54.1%)	51 (45.9%)	9 (8.1%)
Interest in reading	42 (4.6%)	28 (66.7%)	14 (33.3%)	14 (33.3%)
Sensitivity	102 (11.1%)	56 (54.9%)	46 (45.1%)	10 (9.8%)
Wish for variety	116 (12.7%)	62 (53.4%)	54 (46.6%)	8 (6.9%)
Wish to analyze	128 (14.0%)	88 (68.8%)	40 (31.3%)	48 (37.5%)
A | Agreeableness				
Altruism	124 (12.2%)	53 (42.7%)	71 (57.3%)	−18 (−14.5%)
Genuineness	131 (12.8%)	59 (45.0%)	72 (55.0%)	−13 (−9.9%)
Readiness to give feedback	108 (10.6%)	36 (33.3%)	72 (66.7%)	−36 (−33.3%)
Low competitiveness	85 (8.3%)	50 (58.8%)	35 (41.2%)	15 (17.6%)
Good faith	131 (12.8%)	31 (23.7%)	100 (76.3%)	−69 (−52.7%)
Integrity	149 (14.6%)	98 (65.8%)	51 (34.2%)	47 (31.5%)
Search for support	183 (17.9%)	61 (33.3%)	122 (66.7%)	−61 (−33.3%)
Appreciation	109 (10.7%)	55 (50.5%)	54 (49.5%)	1 (0.9%)
C | Conscientiousness				
Task planning	117 (10.3%)	51 (43.6%)	66 (56.4%)	−15 (−12.8%)
Persistence	151 (13.3%)	95 (62.9%)	56 (37.1%)	39 (25.8%)
Dominance	156 (13.7%)	46 (29.5%)	110 (70.5%)	−64 (−41.0%)

(Continued on next page)

Table 2. (Continued)

Facets	Response distribution (%)			
	Overall	Most	Least	Δ
Orderliness	122 (10.7%)	47 (38.5%)	75 (61.5%)	−28 (−23.0%)
Productivity	96 (8.4%)	68 (70.8%)	28 (29.2%)	40 (41.7%)
Self-discipline	117 (10.3%)	82 (70.1%)	35 (29.9%)	47 (40.2%)
Carefulness	103 (9.1%)	59 (57.3%)	44 (42.7%)	15 (14.6%)
Wish to work to capacity	152 (13.4%)	32 (21.1%)	120 (78.9%)	−88 (−57.9%)
Goal orientation	123 (10.8%)	49 (39.8%)	74 (60.2%)	−25 (−20.3%)

Table 3. Response distribution full sample – Big Five domain level

Big Five domain	Response distribution (%)			
	Overall	Most	Least	Δ
N | Emotional Stability	10,450 (19.1%)	4,398 (42.1%)	6,025 (57.9%)	−1,654 (−15.8%)
E | Extraversion	10,630 (19.4%)	4,924 (46.3%)	5,706 (53.7%)	−783 (−7.4%)
O | Openness	10,847 (19.8%)	5,520 (50.9%)	5,327 (49.1%)	193 (1.8%)
A | Agreeableness	11,719 (21.4%)	6,592 (56.3%)	1,465 (43.7%)	1,465 (12.5%)
C | Conscientiousness	11,054 (20.2%)	5,916 (53.5%)	5,138 (46.5%)	778 (7.0%)

here which encourages to continue with the TIRT calibration of the questionnaire based on the collected sample.

Additionally, and for comparison, the response distributions for the Big Five domain level were also calculated for the final sample as depicted in Table 3.

In direct comparison with Table 1, it can be seen that the response allocations are comparable in the 19%–21% range and within the expected even split. Also, *most like me* and *least like me* response options endorsements were stable and even, ranging from 42.1% to 57.9%. Response distributions on facet level are within the expected range as well and comparable to the pilot phase distributions reported above. The detailed facet level results however already heavily represent the sample's personality trait standings and are therefore less of an indicator of the item quality itself can be found in online supplementary material.

Structural Validity Evidence

The hierarchical model consisting of the Big Five domains with their total of 42 facets, the full SEM structure can be seen in Figure 1, was tested using the full sample ($N = 547$) based on the syntax derived from an Excel sheet automatization provided by Brown and Maydeu-Olivares (2011). Run on Mplus (Muthén & Muthén, 1998–2019, version 7.4.), the model converged normally and factor score estimates for each individual could be derived. Correlations and loadings of items and facets were checked manually, a process common in IRT modeling to investigate the adequacy of the model and patterned as expected in magnitude and direction (in most cases, only 10% of the loadings deviated from the expected keyed direction according to the model), meaning high loadings and keyed in the expected direction of their respective facets. The convergence of the model as well as the theory conforming loadings and intercorrelations are therefore seen as evidence of structural validity, despite the unavailability of goodness-of-fit statistics due to model complexity and computational limitations, as outlined above.

Further analyses of structural validity are based on factor score estimates derived from the Mplus TIRT calibration, which were used to gather evidence for structural validity. First, these analyses were conducted separately for every Big Five domain and then forged into a combined overall model. To achieve sophisticated goodness of fit for the separate and overall CFA models, slight modifications to each model were necessary. Those were based on modification indicators derived in the way outlined by Saris et al. (2009) and Bentler and Chou (1992). To achieve satisfactory model fit statistics for each of the Big Five domain measurement models, a total of six out of 42 facets of the B5PS-FC had to be dropped from their respective domain models due to insufficient loadings (maximum of two per domain). Additional constraints, allowing intercorrelations between certain within-domain facets, were implemented as well according to recommendations derived from

Table 4. SEM model modifications for each Big Five dimension

Dimension	Dropped facets	Correlated facets
N	Self-attention (N7)	—
E	Energy (E9)	E1 Sociability ~~ E2 Readiness to take risks
	Forcefulness (E5)	E2 Readiness to take risks ~~ E3 wish for affiliation
		E2 Readiness to take risks ~~ E6 communicativeness
O	Interest in reading (O4)	O3 open-mindedness ~~ O6 wish to analyze
	Creativity (O1)	O5 artistic interest ~~ O9 intellect
		O7 willingness to learn ~~ O8 sensitivity
		O2 wish for variety ~~ O6 wish to analyze
		O2 wish for variety ~~ O8 sensitivity
A	Readiness to give feedback (A4)	A6 good faith ~~ A7 genuineness
		A2 integrity ~~ A5 search for support
		A1 appreciation ~~ A7 genuineness
C	—	C6 carefulness ~~ C8 wish to work to capacity
		C1 dominance ~~ C6 carefulness

Table 5. SEM model fits for the B5PS-FC

Big Five domains	χ^2	*df*	*p*	CFI	RMSEA	SRMR	Ω_w
N | Emotional Stability	80.366	9	<.001	.961	.120	.033	.90
E | Extraversion	175.436	11	<.001	.910	.165	.091	.88
O | Openness	127.116	9	<.001	.937	.155	.087	.79
A | Agreeableness	335.099	11	<.001	.797	.232	.101	.84
C | Conscientiousness	471.762	25	<.001	.813	.181	.082	.88
Overall model	2,606.255	571	<.001	.805	.081	.098	—

Table 6. Big Five domain intercorrelations

Big Five domains	Emotional stability	Extraversion	Openness	Agreeableness	Conscientiousness
Emotional Stability	—				
Extraversion	.01 [.10, .07]	—			
Openness	.15 [.22, .08]	.11 [.05, .18]	—		
Agreeableness	.24 [.32, .15]	.08 [.02, .10]	.07 [.01, .13]	—	
Conscientiousness	.21 [.30, .12]	.34 [.42, .27]	.21 [.28, .13]	.09 [.15, .02]	—

modification indices (Bentler & Chou, 1992; Saris et al., 2009). Adaptations were based on lack of substantial factor loading or negative within-domain correlations and implemented iteratively until satisfactory fit was observed. The implemented adaptations are depicted in Table 4.

The results for goodness-of-fit statistics of the domain and overall models, including the modifications outlined above, are shown in Table 5. Worth noticing are also the construct reliability estimates (weighted Ω) ranging from .79 to .90 in comparison to the B5PS-RS reliability estimates ranging from .66 to .88.

Further analyses on domain level correlation derived from the overall SEM model showed low intercorrelations between the Big Five domains as can be seen in Table 6.

Visualization of the overall model is shown in Figure 2.

Additionally, measurement invariance was checked across different subgroups. The results were interpreted according to the cut-off recommendations for analyses of measurement invariance by Chen (2007) with adequate sample sizes of $N > 300$, stating that noninvariance has to be assumed in cases of $\Delta\text{CFI} \geq -.010$ and $\Delta\text{RMSEA} \geq .015$ or $\Delta\text{SRMR} \geq .010$. At this stage, assumptions for scalar measurement invariance held across age, gender, and

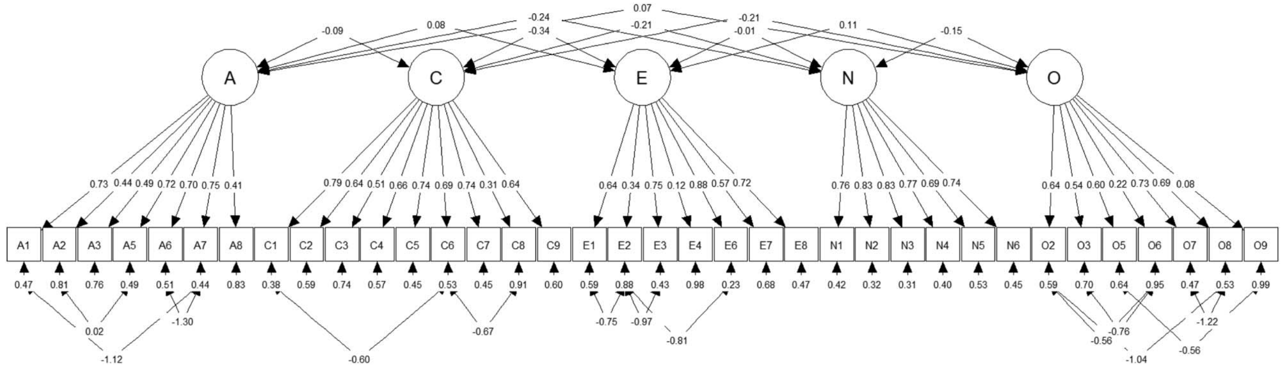


Figure 2. SEM B5PS-FC overall model.

Table 7. Measurement invariance – age, gender, and educational level

Age	χ^2	*df*	*p*	CFI	RMSEA	SRMR
I | Configural MI	4,622.707	2,284	<.001	.783	.087	.112
II | Metric MI	4,694.734	2,377	<.001	.785	.085	.115
III | Scalar MI	4,865.094	2,485	<.001	.779	.084	.117
				ΔCFI	ΔRMSEA	ΔSRMR
$RMSEA_D$ = .000	Delta I | Configural versus II| Metric			.002	–.002	.003
$RMSEA_D$ = .033	Delta II | Metric versus III| Scalar			–.006	–.001	.002
Gender	χ^2	*df*	*p*	CFI	RMSEA	SRMR
I | Configural MI	3,244.324	1,142	<.001	.797	.082	.100
II | Metric MI	3,283.422	1,173	<.001	.796	.081	.101
III | Scalar MI	3,409.505	1,209	<.001	.788	.082	.106
				ΔCFI	ΔRMSEA	ΔSRMR
$RMSEA_D$ = .017	Delta I | Configural versus II| Metric			–.001	–.001	.001
$RMSEA_D$ = .068	Delta II | Metric versus III| Scalar			.008	.001	.005
Education	χ^2	*df*	*p*	CFI	RMSEA	SRMR
I | Configural MI	3,909.413	1713	<.001	.794	.084	.108
II | Metric MI	3,980.666	1775	<.001	.793	.083	.110
III | Scalar MI	4,049.712	1847	<.001	.793	.081	.110
				ΔCFI	ΔRMSEA	ΔSRMR
$RMSEA_D$ = .013	Delta I | Configural versus II| Metric			–.001	–.001	.002
$RMSEA_D$ = .000	Delta II | Metric versus III| Scalar			.000	–.002	.000

educational level as can be seen in Table 7. Since the sole use of changes in CFI, RMSEA, and SRMR for the assessment of MI has been critically reflected in recent studies and additional indicators like $RMSEA_D$ have been proposed (Savalei et al., 2023), we analyzed those as well and cross-checked for all nested comparisons. The results for $RMSEA_D$ confirmed the assessment of MI for age, gender, and education reported here with no $RMSEA_D$ value getting close to the recommended cut-off value of $RMSEA_D < .08$ at which invariance can be stated ($RMSEA_D$ ranging from .000 to .068 across all nested comparisons).

Convergent and Discriminant Validity Evidence

Big Five domain level correlations with scores from the BFSI/S2 were estimated, and the results can be seen in Table 8. Convergent correlations are highest (with the

Table 8. Big Five domain level correlations B5PS-FC versus BFSI/S2

Big Five domain level	N | BFSI	E | BFSI	O | BFSI	A | BFSI	C | BFSI
N | B5PS-FC	**.44** [.37, .50]	.01 [.10, .07]	.03 [.11, .05]	.21 [.29, .13]	.02 [.11, .06]
E | B5PS-FC	.07 [.02, .15]	**.29** [.21, .36]	.15 [.07, .23]	.10 [.01, .18]	.21 [.29, .13]
O | B5PS-FC	.04 [.05, .12]	.18 [.10, .26]	**.24** [.16, .32]	.18 [.10, .26]	.04 [.13, .04]
A | B5PS-FC	.16 [.24, .08]	.13 [.21, .05]	.12 [.21, .04]	**.16** [.08, .24]	.18 [.26, .10]
C | B5PS-FC	.27 [.35, .19]	.19 [.27, .11]	.24 [.32, .16]	.11 [.19, .03]	**.27** [.19, .34]

Note. Convergent correlations in bold, discriminant correlations in off-diagonal cells.

exception of Agreeableness) while discriminant correlations are mostly lower or even negative. Overall, keyed direction and magnitude turned out in expected ranges.

Discussion

This study aimed at adapting an existing Big Five rating scale measure (B5PS-RS) to an MFC format. The existing measure used situation vignettes to contextualize the assessment. This feature was kept in the new version. The advantage of the MFC format is in preventing typical response distortions. A Thurstonian IRT (Brown & Maydeu-Olivares, 2011) modeling approach was applied to leverage its ability to derive nonipsative results that usually result from sum scoring MFC questionnaire response data, allowing interpersonal comparisons based on MFC test results. Main goal was to deliver a proof of concept of a situation-based, multifaceted Big Five personality inventory. During the adaptation process, insights into MFC design were gained which will be outlined below. Moreover, empirical evidence supporting the reliability, structural, and construct validity of the scores derived from the MFC measure were obtained.

Insights Into MFC Questionnaire Design

Designing and constructing MFC questionnaires is accompanied by increased efforts in test construction due to the peculiarities of the context-wise item interdependencies. All items within a block are directly related to each other, and in the current endeavor, this issue was intertwined with the placing of item blocks within a situation vignette. The difficulty here is to find items reflecting different Big five domains which are equally realistic responses to the vignette. While this challenge demanded more time than just pairing items by social desirability, we hypothesize that there might be an advantage with regard to fakability. In particular, by contextualizing each item block and presenting options equally desirable in a given situation, faking should be made even more difficult. Future research needs to test this hypothesis by comparing the fakability of the B5PS-MFC with other MFC measures.

Our experiences with designing situation-based item blocks leads us to strongly advocate the use of several pilot phases in MFC item design, both qualitative and quantitative, as demonstrated in this test construction process. The success of our iterative item adaptation process can be derived from the response distributions reported here. Leveraging both experts and practitioners' knowledge prior to quantitative pilot phasing has proved to be invaluable to the overall item quality. The first pilot versions of the questionnaire would have not been able to achieve such evenly distributed response allocations. Matching an almost perfect 20% split between Big Five dimension endorsements and carrying this distribution over to the final sample can be mostly ascribed to this iterative process. Even on facet level, the item endorsement spreads mostly evenly, only slightly shifting toward one of the two response options in some facets which might just reflect the actual trait standings in the sample. With this we are stressing the need for several pilot phases, both qualitative and quantitative before going all in with data collection for TIRT calibration. Many issues later arising can be prevented beforehand with a thorough iterative item optimization process.

Another aspect to point out from our experience is also the increased effort required by the test taker to work on the MFC questionnaire. Processing several binary comparisons in an MFC block turns out to be a more challenging task compared to single statement ratings known from rating scale response formats. An observation already mentioned by Brown and Maydeu-Olivares (2012) and Sass et al. (2020), the increased cognitive strain has been reported by many test takers in this study. Consequently, this should be taken into account when designing MFC questionnaires, especially considering overall questionnaire length, item count within each block, and complexity of item content itself. In our effort to construct a situation-based and contextualized MFC questionnaire featuring four items per block, we surely pushed the boundaries of what test takers are used to and able to handle. It is of relevance here that the sample used for the TIRT calibration in this study mostly consists of people with higher education while the questionnaire's contextualized situation vignettes are based in white-collar work environments in organizational settings. This combination somewhat limits the applicability of the questionnaire outside of its target population. Both language proficiency and cognitive competence will impose a restriction on the usage in situations where both the test takers understanding of and the familiarity with the subject matter are limited. First, because lower educational levels are not represented in the calibration sample, and second, because familiarity with organizational settings and the capability to deal with more complex binary comparisons is vital to a proper test completion.

However, an important aspect that was reported by our test takers as well is that MFC response formats initiate a process of self-reflection for most test takers to really explore their own ranking of the items within a block, an interesting side effect that could also contribute to the validity of the test score interpretations. Moreover, it is important to stress here that the intended target population should possess sufficient cognitive abilities to handle the complexity. For the B5PS-FC, this is assured as long as test takers are part of the intended target population.

Psychometric Evaluation of the B5PS-FC

Apart from the qualitative and descriptive indicators of a successful test adaptation discussed above, the psychometric evaluation of the scores derived from the B5PS-FC also substantiates the proof of concept of a situation-based MFC questionnaire construction. Despite the lack of goodness-of-fit statistics, which are unavailable with current computational power, the item loading checks and follow-up CFA/SEM analyses showed satisfactory levels of model fit to assume a successful response format adaptation and an intact construct structure in the B5PS-FC (Table 5).

Indicators of structural validity can be found in the Mplus TIRT model convergence and second in follow-up SEM analyses based on factor scores derived from the TIRT calibration. Acceptable fit indices for each Big Five dimension and the overall model (Table 5) demonstrate how the psychometric quality of the hierarchical, large-scale model represented in the data of the B5PS-FC. Especially considering that large and complex personality inventories usually perform poorly in CFAs (Hopwood & Donnellan, 2010), the results found here are strong indicators of structural validity. Adding to that are high reliability estimates (Ω_w ranging from .79 to .90), actually outperforming estimates for the B5PS-RS.

Important for the eventually intended application of the questionnaire, analyses of measurement invariance across age, gender, and educational level support the assumption of scalar MI (Table 7). Additionally, low intercorrelations between Big Five domains in the B5PS-FC (Table 8) and the expectedly small to moderate convergent correlations with scores from a standard Big Five rating scale questionnaire (BFSI/S2) were found. The small to moderate correlations found here were expected since the BFSI/S2 features a noncontextualized item design and a different facet structure than the B5PS. Moreover, the findings are in line with prior research showing that there is a gap between operationalizations in the Costa and McCrae tradition (2008) which was used in the BFSI and the Goldberg tradition used in the B5PS (Miller et al., 2011).

Apart from that, a more tangible issue arises when investigating correlated facets in the B5PS-FC, especially in the Openness domain, as depicted in Table 4. The correlated facets are mostly congruent with those showing suspicious item loadings. First analyses were pointing into the direction of the aspects of Big Five facet being correlated here. Since all facets of the B5PS can be represented by two aspects, comprising two characteristics of the same underlying feature (DeYoung et al., 2007; Ziegler et al., 2019), those might be the cause of correlations found in the SEM analyses. However, this is not an exhaustive explanation since it does not apply to all correlations between facets. Further analyses are needed here to explore the issues that could potentially have been fixed with another iterative item design process step to further optimize item content. Still, considering the accumulated evidence, use of the B5PS-FC for research purposes, especially for the domain scores, seems feasible.

Overall, the psychometric evaluation implies that adapting the response format of the B5PS-RS to the B5PS-FC was managed without substantially changing (the structure of) the construct to be measured and keeping it intact in measures of structural validity. In further steps, test criterion validity evidence is needed to substantiate these findings.

Limitations of the Present Study

The adaptation process reported in this article has some limitations. For one, technical or computational hardware availability limited the Mplus-based TIRT calculations for goodness of fit and standard error estimations. These were adhered to by manual checks of item loadings and follow-up confirmatory factor analyses based on factor scores derived from the TIRT calibration. Both item loadings and the implemented modifications for the structure equation modeling could be used for further optimization of items in the future.

A second limitation might be that the assumed intercorrelations for within-dimension facets used in the TIRT modeling and the additional assumption of noncorrelated Big Five domains are strong ones, but they are based on empirical data available from the B5PS-RS (Ziegler et al., 2019), theoretical implications derived from Big Five models used (Miller et al., 2011), and supported by the inventors of the TIRT modeling approach (Brown & Maydeu-Olivares, 2012). The necessary modifications to the SEM analyses that were mostly based on empirical modification indicators are another topic of concern. However, we assume that an overall larger sample size and some minor item adjustments would be able to solve these issues. Besides that, a replication of the (modified) models in a second sample would be necessary as well and should be undertaken in the future.

However, there is research showing that the nomological net of MFC and RS versions of the same questionnaire is comparable (Walton et al., 2019), and this was corroborated by the current findings. Still, further research needs to provide more evidence to support this assumption.

Conclusion and Outlook

We successfully delivered the proof of concept to a situation-based Big Five inventory featuring an MFC response format, calibrated according to the TIRT modeling approach to derive nonipsative data, suitable for interpersonal comparison of test scores. All this was achieved while still maintaining a hierarchical dimension facet structure based on an already existing rating scale-based personality questionnaire (B5PS-RS). This allows us to continue the exploration of the MFC format in direct comparison to its rating scale version, to take a closer look at construct comparability, and to tackle the topic of social desirable responding (SDR) or faking mitigation in low- and high-stakes assessment situations. Only then will it be fair to judge whether the B5PS-FC truly outperforms its rating scale version, though we hope to establish MFC formats as a viable alternative to rating scale formats in the future. The foundations have been laid.

References

Arendasy, M. (2011). *Big-Five Struktur Inventar (BFSI)* [Big Five Structure Inventory]. Schuhfried GmbH.

Bäckström, M., & Björklund, F. (2014). Social desirability in personality inventories: The nature of the evaluative factor. *Journal of Individual Differences*, *35*(3), 144–157. https://doi.org/10.1027/1614-0001/a000138

Baron, H. (1996). Strengths and limitations of ipsative measurement. *Journal of Occupational and Organizational Psychology*, *69*(1), 49–56. https://doi.org/10.1111/j.2044-8325.1996.tb00599.x

Bentler, P. M., & Chou, C.-P. (1992). Some new covariance structure model improvement statistics. *Sociological Methods & Research*, *21*(2), 259–282. https://doi.org/10.1177/0049124192021002006

Brown, A. (2010). *How item response theory can solve problems of ipsative data* [Doctoral dissertation, Universitat de Barcelona]. https://kar.kent.ac.uk/44768/

Brown, A. (2015). Personality assessment, forced-choice. In J. Wright (Ed.), *International Encyclopedia of the social and behavioural sciences* (2nd ed., pp. 840–848). Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.25084-8

Brown, A. (2016). Item response models for forced-choice questionnaires: A common framework. *Psychometrika*, *81*(1), 135–160. https://doi.org/10.1007/s11336-014-9434-9

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement*, *71*(3), 460–502. https://doi.org/10.1177/0013164410375112

Brown, A., & Maydeu-Olivares, A. (2012). Fitting a Thurstonian IRT model to forced-choice data using Mplus. *Behavior Research Methods*, *44*(4), 1135–1147. https://doi.org/10.3758/s13428-012-0217-x

Brown, A., & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods*, *18*(1), 36–52. https://doi.org/10.1037/a0030641

Brown, A., & Maydeu-Olivares, A. (2018). Modeling forced-choice response formats. In P. Irwing, T. Booth, & D. Hughes (Eds.), *The Wiley handbook of psychometric testing* (pp. 523–570). Wiley-Blackwell.

Bürkner, P.-C., Schulte, N., & Holling, H. (2019). On the statistical and practical limitations of Thurstonian IRT models. *Educational and Psychological Measurement*, *79*(5), 827–854. https://doi.org/10.1177/0013164419832063

Cao, M., & Drasgow, F. (2019). Does forcing reduce faking? A meta-analytic review of forced-choice personality measures in high-stakes situations. *The Journal of applied psychology*, *104*(11), 1347–1368. https://doi.org/10.1037/apl0000414

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, *14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Costa, P. T. Jr., & McCrae, R. R. (2008). The Revised NEO Personality Inventory (NEO-PI-R). In G. J. Boyle, G. Matthews, & D. H. Saklofske (Eds.), *The SAGE handbook of personality theory and assessment. Personality measurement and testing* (Vol. 2, pp. 179–198). Sage Publications, Inc. https://doi.org/10.4135/9781849200479.n9

DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the Big Five. *Journal of Personality and Social Psychology*, *93*(5), 880–896. https://doi.org/10.1037/0022-3514.93.5.880

Dueber, D. M., Love, A. M. A., Toland, M. D., & Turner, T. A. (2019). Comparison of single-response format and forced-choice format instruments using Thurstonian item response theory.

Educational and Psychological Measurement, *79*(1), 108–128. https://doi.org/10.1177/0013164417752782

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Frick, S., Brown, A., & Wetzel, E. (2023). Investigating the normativity of trait estimates from multidimensional forced-choice data. *Multivariate Behavioral Research*, *58*(1), 1–29. https://doi.org/10.1080/00273171.2021.1938960

Fuechtenhans, M., & Brown, A. (2022). How do applicants fake? A response process model of faking on multidimensional forced-choice personality assessments. *International Journal of Selection and Assessment*, *31*(1), 105–119. https://doi.org/10.1111/ijsa.12409

Goldberg, L. R. (1990). An alternative “description of personality”: The Big-Five factor structure. *Journal of Personality and Social Psychology*, *59*(6), 1216–1229. https://doi.org/10.1037/0022-3514.59.6.1216

Greiff, S., & Heene, M. (2017). Why psychological assessment needs to start worrying about model fit [Editorial]. *European Journal of Psychological Assessment*, *33*(5), 313–317. https://doi.org/10.1027/1015-5759/a000450

Hallquist, M. N., & Wiley, J. F. (2018). Mplus Automation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(4), 621–638. https://doi.org/10.1080/10705511.2017.1402334

Heene, M., Hilbert, S., Draxler, C., Ziegler, M., & Bühner, M. (2011). Masking misfit in confirmatory factor analysis by increasing unique variances: A cautionary note on the usefulness of cutoff values of fit indices. *Psychological Methods*, *16*(3), 319–336. https://doi.org/10.1037/a0024917

Henry, L., & Wickham, H. (2022). *purrr: Functional programming tools*. http://purrr.tidyverse.org

Hopwood, C. J., & Donnellan, M. B. (2010). How should the internal structure of personality inventories be evaluated? *Personality and Social Psychology Review*, *14*(3), 332–346. https://doi.org/10.1177/1088868310361240

Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Kemper, C. J., Trapp, S., Kathmann, N., Samuel, D. B., & Ziegler, M. (2019). Short versus long scales in clinical assessment: Exploring the trade-off between resources saved and psychometric quality lost using two measures of obsessive–compulsive symptoms. *Assessment*, *26*(5), 767–782. https://doi.org/10.1177/1073191118810057

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research*, *44*(3), 486–507. https://doi.org/10.1177/0049124114543236

Meade, A. (2004). Psychometric problems and issues involved with creating and using ipsative measures for selection. *Journal of Occupational and Organizational Psychology*, *77*(4), 531–551. https://doi.org/10.1348/0963179042596504

Miller, J. D., Gaughan, E. T., Maples, J., & Price, J. (2011). A comparison of agreeableness scores from the Big Five Inventory and the NEO PI-R: Consequences for the study of narcissism and psychopathy. *Assessment*, *18*(3), 335–339. https://doi.org/10.1177/1073191111411671

Muthén, L. K., & Muthén, B. O. (1998–2019). *Mplus* [Computer software]. https://www.statmodel.com

Pace, V. L., & Brannick, M. T. (2010). How similar are personality scales of the “same” construct? A meta-analytic investigation. *Personality and Individual Differences*, *49*(7), 669–676. https://doi.org/10.1016/j.paid.2010.06.014

Pavlov, G., Shi, D., Maydeu-Olivares, A., & Fairchild, A. (2021). Item desirability matching in forced-choice test construction. *Personality and Individual Differences*, *183*. Article 111114. https://doi.org/10.1016/j.paid.2021.111114

R Core Team. (2021). *R: A language and environment for statistical computing* (Version 4.0.4) [Computer software]. R Foundation for Statistical Computing. https://www.Rproject.org/

Revelle, W. (2019). *psych: Procedures for personality and psychological research* (Version 1.8.12) [Computer software]. https://CRAN.R-project.org/package=psych

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Rouco, V., Cengia, A., Roberts, R., Kemper, C., & Ziegler, M. (2022). The Berlin Multi-Facet Personality Inventory. *Psychological Test Adaption and Development*, *3*, 23–34. https://doi.org/10.1027/2698-1866/a000021

RStudio Team. (2021). *RStudio: Integrated development for R* (Version 1.4.1106) [Computer software]. http://www.rstudio.com

Saris, W. E., Satorra, A., & van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling*, *16*(4), 561–582. https://doi.org/10.1080/10705510903203433

Sass, R., Frick, S., Reips, U.-D., & Wetzel, E. (2020). Taking the test taker’s perspective: Response process and test motivation in multidimensional forced-choice versus rating scale instruments. *Assessment*, *27*(3), 572–584. https://doi.org/10.1177/1073191118762049

Savalei, V., Brace, J. C., & Fouladi, R. T. (2023). We need to change how we compute RMSEA for nested model comparisons in structural equation modeling. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000537

Schulte, N., Holling, H., & Bürkner, P.-C. (2021). Can high-dimensional questionnaires resolve the ipsativity issue of forced-choice response formats? *Educational and Psychological Measurement*, *81*(2), 262–289. https://doi.org/10.1177/0013164420934861

Shaffer, J. A., & Postlethwaite, B. E. (2012). A matter of context: A meta-analytic investigation of the relative validity of contextualized and noncontextualized personality measures. *Personnel Psychology*, *65*(3), 445–493. https://doi.org/10.1111/j.1744-6570.2012.01250.x

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multi-unidimensional pairwise-preference model. *Applied Psychological Measurement*, *29*(3), 184–203. https://doi.org/10.1177/0146621604273988

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Walton, K. E., Cherkasova, L., & Roberts, R. D. (2020). On the validity of forced choice scores derived from the Thurstonian item response theory model. *Assessment*, *27*(4), 706–718. https://doi.org/10.1177/1073191119843585

Wetzel, E., Böhnke, J. R., & Brown, A. (2016). Response biases. In F. T. L. Leong, D. Bartram, F. Cheung, K. F. Geisinger, & D. Iliescu (Eds.), *The ITC international handbook of testing and assessment* (pp. 394–363). Oxford University Press. https://doi.org/10.1093/med:psych/9780199356942.003.0024

Wetzel, E., & Frick, S. (2020). Comparing the validity of trait estimates from the multidimensional forced-choice format and the rating scale format. *Psychological Assessment*, *32*(3), 239–253. https://doi.org/10.1037/pas0000781

Wetzel, E., Frick, S., & Brown, A. (2021). Does multidimensional forced-choice prevent faking? Comparing the susceptibility of the multidimensional forced-choice format and the rating scale format to faking. *Psychological Assessment*, *33*(2), 156–170. https://doi.org/10.1037/pas0000971

Wickham, H., François, R., Henry, L., & Müller, K. (2022). *dplyr: A Grammar of data manipulation*. https://dplyr.tidyverse.org

Xie, Y. (2022). *knitr: A general-purpose package for dynamic report generation in R*. https://yihui.org/knitr/

Ziegler, M. (2014). *B5PS. Big Five Inventory of personality in occupational situations*. Schuhfried GmbH.

Ziegler, M., & Bühner, M. (2009). Modeling socially desirable responding and its effects. *Educational and Psychological Measurement*, *69*(4), 548–565. https://doi.org/10.1177/0013164408324469

Ziegler, M., Horstmann, K. T., & Ziegler, J. (2019). Personality in situations: Going beyond the OCEAN and introducing the Situation Five. *Psychological Assessment*, *31*(4), 567–580. https://doi.org/10.1037/pas0000654

Ziegler, M., & Kemper, C. (2013). Extreme response style and faking: Two sides of the same coin? In P. Winker, R. Porst, & N. Menold (Eds.), *Interviewers' deviations in surveys: Impact, reasons, detection and prevention* (pp. 217–233). Peter Lang. https://orbilu.uni.lu/handle/10993/21205

History
Received October 15, 2022
Revision received March 30, 2023
Accepted May 15, 2023
Published online August 8, 2023

Section: Personality

Acknowledgments
The authors wish to thank all participants for their valuable contributions to this study.

Conflict of Interest
The authors report no conflict of interest.

Publication Ethics
Informed consent was obtained from all participants included in the study.

Authorship
Alexander Leonard Schünemann: conceptualization, formal analyses, investigation, methodology, writing – original draft; Matthias Ziegler: conceptualization, methodology, supervision, writing – review and editing. All authors approved the final version of the article.

Open Science
The data that support the findings of this study are available on request from the corresponding author, Alexander Leonard Schünemann. The data are not publicly available due to privacy restrictions.

The online supplementary materials are available at https://osf.io/hfws2/

ORCID
Alexander Leonard Schünemann
https://orcid.org/0000-0002-0608-7202
Matthias Ziegler
https://orcid.org/0000-0003-4994-9519

Alexander Leonard Schünemann
Institute of Psychology
Humboldt-Universität zu Berlin
Rudower Chaussee 19
12489 Berlin
Germany
leonard.schuenemann@hu-berlin.de