Editorial

The Multidimensional Forced-Choice Format as an Alternative for Rating Scales – Current State of the Research

Eunike Wetzel,[1] Susanne Frick,[2] and Samuel Greiff[3]

[1] Department of Psychology, Otto-von-Guericke University Magdeburg, Germany
[2] Department of Psychology, University of Mannheim, Germany
[3] Institute of Cognitive Science and Assessment, University of Luxembourg, Luxembourg

When constructing a questionnaire to assess a psychological construct, one important decision researchers have to make is how to collect responses from test takers; that is, which response format to implement. We argued in a previous editorial published in *European Journal of Psychological Assessment (EJPA)* that this decision deserves more attention and should be an explicit step in the test construction process (Wetzel & Greiff, 2018). The reason for this is that it can be a consequential decision that influences the validity of conclusions we draw about test takers' trait levels or about relations between constructs and criteria (Brown & Maydeu-Olivares, 2013; Wetzel & Frick, 2020). In this editorial, which can be considered a follow-up to the first one, we will take a closer look at two response formats[1]: rating scales (RS), the current default in most questionnaires, and the multidimensional forced-choice (MFC) format, an alternative that is currently the focus of a considerable body of research. We will first define the two formats and point out some of their advantages and disadvantages. Then, we will provide a summary and evaluation of research comparing RS and MFC. Third, we will draw some preliminary conclusions on the feasibility of applying MFC as an alternative to RS. Fourth, we will point out some open research questions. We will end with some recommendations and implications for readers and authors of *EJPA*. In this editorial, the overall goal is to give researchers and test users an overview of the current state of the research on RS versus MFC and to provide guidance on the feasibility of applying MFC in research on psychological assessment.

Rating Scales

Most available questionnaires collect item responses by asking test takers to describe themselves on RS such as *strongly disagree – disagree – neutral – agree – strongly agree*. Importantly, each item is rated individually and – at least in theory – independently from the other items measuring the same trait. RS overall allow reliable and valid assessments, especially with homogeneous samples. Nevertheless, there are numerous criticisms of them, for example, that they can be faked easily and that participants differ in how they interpret and use the response categories (e.g., Hernández et al., 2004; Holden & Book, 2011; Krosnick, 1999; Wetzel et al., 2016). Test construction for RS instruments is fairly straightforward with guidelines being described for example in Simms (2008).

Multidimensional Forced-Choice Format

In the forced-choice format, two or more items are presented to participants simultaneously in a block. In the case of the multidimensional version of the forced-choice format, which we focus on here, these items measure different traits. Responses are given directly with the items and different instructions exist. For example, test takers could be asked to select the item that is most like them and least like them or they could be asked to rank the items according to how well the items describe them. Figure 1 shows different variants of the MFC format (pairs, triplets, quads) with different instructions. For an excellent and more thorough introduction to the MFC format see Brown and Maydeu-Olivares (2018a). The MFC format eliminates any response biases that are specific to RS such as extreme response style or acquiescence. It also eliminates rater biases such as halo effects and severity/leniency, though other response biases (faking, careless responding) can still occur. Test construction with the MFC format is more complex than with RS because it requires a number of additional considerations.

[1] The multidimensional forced-choice format is both an item and a response format. For simplicity in the comparison with rating scales, we refer to it as response format.

https://doi.org/10.1027/1015-5759/a000609

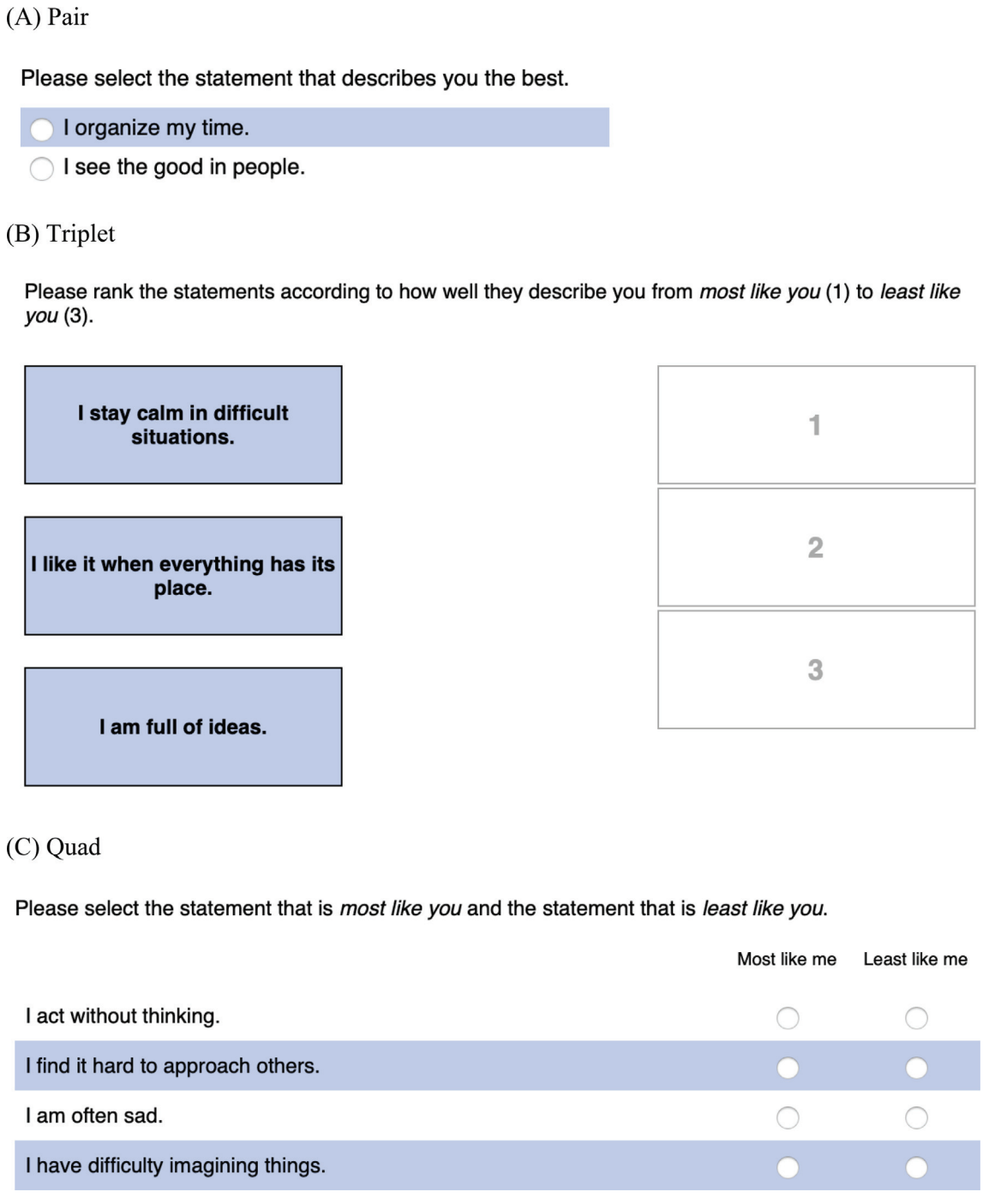

Figure 1. Examples for the multidimensional forced-choice (MFC) format.

In particular, the decision of how to combine items to blocks is important because it has implications for the recovery of absolute trait levels and for respondents' ability to fake. To reduce faking, items within blocks can be matched by desirability. In the MFC format, more items or larger item blocks (e.g., quads instead of triplets) are needed to achieve similar levels of reliability as with RS. For test construction guidelines on MFC formats see Brown and Maydeu-Olivares (2018a).

Data Analysis of MFC Data

The process of responding to MFC item blocks is overall similar to that of responding to RS items, but additionally requires participants to weigh the items within a block against each other (Sass et al., 2020). Thus, in the MFC format, respondents provide comparative judgments, whereas with RS they provide absolute judgments. This has important implications for data analysis. Using classical scoring

methods that simply assign scores to ratings is appropriate to obtain respondents' absolute trait levels (i.e., normative data) with RS. In contrast, with MFC data, classical scoring leads to ipsative data, which is only appropriate for intraindividual comparisons and not for interindividual comparisons or any analyses based on correlations (Hicks, 1970). Thus, classically scored MFC data cannot be used for common assessment contexts such as selection decisions or analyses used commonly in research such as factor analysis, correlations, or the computation of reliability indices. Ipsativity/normativity can be seen as a continuum and one way to reduce the degree of ipsativity is to include negatively keyed items, resulting in partially ipsative (but still not normative) data. Absolute trait levels can be derived from MFC data by applying an item response model such as the Thurstonian item response model (Brown & Maydeu-Olivares, 2011). However, when the testing purpose is to obtain an individual's profile (e.g., of their levels of interest in different areas), classical scoring is also appropriate with MFC data. For more information on scoring MFC data and the properties of ipsative or partially ipsative data, see Brown and Maydeu-Olivares (2018a) and Hicks (1970). For an overview over different item response theory approaches to analyzing MFC data and their integration into a common framework, see Brown (2016).

Research Comparing RS and MFC

Research comparing RS and MFC has focused mainly on two issues: validity and faking. Recent studies applying normative scoring for both formats mostly found similar criterion-related validities for RS and MFC (Lee et al., 2018; Wetzel & Frick, 2020; Zhang et al., 2019), whereas older studies applying ipsative or partially ipsative scoring for MFC often found higher criterion-related validities for MFC (Bartram, 2007; Salgado & Táuriz, 2014). Besides the differences in scoring methods, possible reasons for the inconsistent findings include that older studies used more objective criteria (e.g., job performance), whereas some newer studies used "criteria" assessed with RS. With respect to construct validity, several studies applying normative scoring for MFC found descriptively similar levels of construct validity between both formats (Brown & Maydeu-Olivares, 2013; Lee et al., 2018; Zhang et al., 2019). One study that tested preregistered hypotheses and included comparisons within formats (i.e., MFC with MFC and RS with RS) found inconclusive results (Wetzel & Frick, 2020). On the one hand, some construct validity coefficients were larger for MFC and others for RS. On the other hand, the correspondence between self-ratings and other-ratings was consistently better for MFC. Thus, the evidence on the validity of MFC versus RS is still inconclusive. Nevertheless, in absolute terms, trait estimates from MFC questionnaires can achieve good construct and criterion-related validity.

The second issue that has stimulated a lot of research is whether MFC questionnaires are fake-proof or at least less susceptible to faking than RS questionnaires. In fact, the idea that MFC would be harder to fake was the main reason it was advocated in the first place. So far, most existing research found that less faking occurs with MFC than RS (e.g., Christiansen et al., 2005; Jackson et al., 2000). Cao and Drasgow (2019) recently meta-analyzed previous research on faking Big Five MFC measures and found an overall effect size of $d = 0.06$ (range from 0.00 for neuroticism and openness to 0.23 for conscientiousness), which is substantially lower than the effect sizes for faking Big Five RS measures (range 0.11–0.45; Birkeland et al., 2006). Cao and Drasgow (2019) also checked scoring method (ipsative, partially ipsative, normative) as a moderator and found significant faking effects only with ipsative and partially ipsative scoring, but not normative scoring of MFC data. Wetzel et al. (2020) confirmed previous findings with a Thurstonian item response model analysis of MFC data and additionally showed that the degree to which MFC can be faked depends on how well the items within blocks are matched in terms of their desirability.

A few other issues in the comparison of MFC versus RS have been the topic of recent investigation. For example, Sass et al. (2020) did not find any differences in respondents' test motivation between groups filling out the same items with RS or different versions of the MFC format. Lee et al. (2019) found that the MFC version of a Big Five questionnaire produced more stable personality profile solutions than the RS version.

Preliminary Conclusions on the Feasibility of MFC as an Alternative to RS

When deciding whether to apply MFC or RS, researchers are faced with a cost-benefit trade-off. Test construction is easier with RS than MFC and test scores from RS instruments are generally more reliable than test scores from MFC instruments of the same length (see above). RS items are evaluated independently, whereas the items in MFC blocks interact and this can influence the response process as well as the items' psychometric properties (Lin & Brown, 2017). To derive absolute trait levels, MFC data have to be analyzed with appropriate item response theory models, whereas for RS data, this is possible even with scoring according to classical test theory. MFC eliminates response styles in self-report questionnaires (e.g., acquiescence) and rater biases (e.g., severity/leniency), whereas these biases occur when questionnaires are administered with RS.

MFC is harder to fake than RS, especially when items within blocks have been matched carefully with respect to their desirability (Wetzel et al., 2020).

Which of these aspects is most relevant will depend on the assessment context and the specific goals of the assessment. For example, if the testing purpose is selection, faking will be the most important concern and the ability of MFC to reduce faking will outweigh the costs. If the assessment is a cross-cultural assessment, in which respondents from different countries are expected to differ in their RS use, MFC might also be preferred. On the other hand, if there are severe constraints on testing time, RS might be preferred to achieve adequate reliability. Or, if there are few resources for the development of the instrument, RS might also be preferred.

Open Research Questions on MFC as an Alternative to RS

There are a number of open questions that we need answers to in order to be able to make clearer recommendations on the feasibility of using the MFC format. For example, previous research compared correlations between traits assessed with MFC and criteria or other constructs with those obtained from RS. Response biases such as response styles potentially influenced both previous empirical findings and comparisons with the MFC format. We do not know what the true correlations are and using correlations from RS as a benchmark against which we compare correlations from MFC may be a mistake. Thus, we need stronger study designs that include within-format comparisons for MFC and criteria that are not assessed with RS. In addition, we need study designs that rely on more than just self-reports to reduce the influence of response styles. Research so far has not considered whether there are any response biases that might be unique to the MFC format. Are there any and if yes, what is their influence? Previous research on validity in many cases analyzed homogeneous samples (e.g., students from one country). Therefore, it is unclear whether findings generalize to more heterogeneous samples in which response biases might exert a stronger influence. This is particularly important with respect to samples with lower cognitive abilities or education level who might be overburdened with forced-choice tasks, especially with item blocks larger than triplets. In addition, the generalizability of research findings to other constructs besides personality traits could also be investigated in future research.

Another open research question is how to match items within blocks with respect to their desirability. Previous research mainly used item means from an administration with RS or desirability ratings of the individual items. Both methods may be inadequate because items within MFC blocks can still differ in their relative desirability. Related to this, another open research question is whether it is easier to fake MFC blocks that contain mixed positively and negatively keyed items compared with MFC blocks that contain only equally keyed items when all items have been matched for desirability. Previous research on faking in MFC versus RS mostly used instructed faking designs. Thus, studies in naturalistic settings (e.g., with applicant samples) are needed.

A few approaches to combining MFC with RS have been developed. Xiao et al. (2017) suggested simply administering several RS items in addition to MFC blocks. This approach improved model convergence and the recovery of latent trait levels in their simulation study. Brown and Maydeu-Olivares (2018b) proposed the graded preference format, in which respondents are faced with a choice between two items, but have to indicate the strength of their preference for one of the items on an RS. Future research could investigate whether these proposals are able to capitalize on the strengths of both formats without being limited by their difficulties.

Implications for Readers and Authors of *EJPA*

Research on MFC is vibrant and productive and we hope that some of the open questions noted above will be addressed in the next years – maybe by *EJPA* authors. We have tried to give readers a balanced synopsis of the current state of the research on MFC as an alternative to RS and to show that more research in this area – an area at the core of *EJPA* – is needed. It has hopefully become clear that both formats have their strengths and challenges. For test constructors, this implies that thinking about which format to use should be an explicit step during test construction and the costs and benefits of the different formats should be weighed against each other (Wetzel & Greiff, 2018). For test users, this implies that the costs and benefits of the different formats should be taken into account during test selection. Keeping in mind that the instrument should be tailored to the testing purpose (Ziegler, 2014), sometimes RS instruments may be more appropriate while sometimes MFC instruments may be the best choice. We hope to see research submitted to *EJPA* that targets some of the open questions mentioned above either from a more assessment-related methodological perspective or through direct applications and comparisons of the different formats for the assessment of specific traits and the measures associated with them. Ultimately, this will likely lead to a broader portfolio of methods as well as measures available to the community of psychological assessment.

References

Bartram, D. (2007). Increasing validity with forced-choice criterion measurement formats. *International Journal of Selection and Assessment, 15*(3), 263–272. https://doi.org/10.1111/j.1468-2389.2007.00386.x.

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment, 14*(4), 317–335. https://doi.org/10.1111/j.1468-2389.2006.00354.x.

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. (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. (2018a). **[Author: add page range]**Modeling of forced-choice response formats. In P. Irwing, T. Booth, & D. Hughes (Eds.), *The Wiley handbook of psychometric testing*. Wiley.

Brown, A., & Maydeu-Olivares, A. (2018b). Ordinal factor analysis of graded-preference questionnaire data. *Structural Equation Modeling, 25*(4), 516–529. https://doi.org/10.1080/10705511.2017.1392247.

Cao, M., & Drasgow, F. (2019). **[Author: add volume and page range]** Does forcing reduce faking? A meta-analytic review of forced-choice personality measures in high-stakes situations. *Journal of Applied Psychology, 104*(11), 1347–1368. **[Author: please check & approve the details added]** https://doi.org/10.1037/apl0000414.

Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering forced-choice item formats for applicant personality assessment. *Human Performance, 18*(3), 267–307. https://doi.org/10.1207/s15327043hup1803_4.

Hernández, A., Drasgow, F., & González-Romá, V. (2004). Investigating the functioning of a middle category by means of a mixed-measurement model. *Journal of Applied Psychology, 89*(4), 687–699. https://doi.org/10.1037/0021-9010.89.4.687.

Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin, 74*(3), 167–184. https://doi.org/10.1037/h0029780.

Holden, R. R., & Book, A. S. (2011). **[Author: add page range]** Faking does distort self-report personality assessment. In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.), *New perspectives on faking in personality assessment*. Oxford University Press.

Jackson, D. N., Wroblewski, V. R., & Ashton, M. C. (2000). The impact of faking on employment tests: Does forced choice offer a solution?. *Human Performance, 13*(4), 371–388. https://doi.org/10.1207/S15327043hup1304_3.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology, 50*, 537–567. https://doi.org/10.1146/annurev.psych.50.1.537.

Lee, P., Joo, S. H., & Lee, S. (2019). Examining stability of personality profile solutions between Likert-type and multidimensional forced choice measure. *Personality and Individual Differences, 142*, 13–20. https://doi.org/10.1016/j.paid.2019.01.022.

Lee, P., Lee, S., & Stark, S. (2018). Examining validity evidence for multidimensional forced choice measures with different scoring approaches. *Personality and Individual Differences, 123*, 229–235. https://doi.org/10.1016/j.paid.2017.11.031.

Lin, Y., & Brown, A. (2017). Influence of context on item parameters in forced-choice personality assessments. *Educational and Psychological Measurement, 77*(3), 389–414. https://doi.org/10.1177/0013164416646162.

Salgado, J. F., & Táuriz, G. (2014). The Five-Factor Model, forced-choice personality inventories and performance: A comprehensive meta-analysis of academic and occupational validity studies. *European Journal of Work and Organizational Psychology, 23*(1), 3–30. https://doi.org/10.1080/1359432x.2012.716198.

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.

Simms, L. J. (2008). Classical and modern methods of psychological scale construction. *Social and Personality Psychology Compass, 2*, 414–433. https://doi.org/10.1111/j.1751-9004.2007.00044.x.

Wetzel, E., Böhnke, J. R., & Brown, A. (2016). Response biases. In F. R. L. Leong, B. Bartram, F. Cheung, K. F. Geisinger, & D. Iliescu (Eds.), *The ITC International handbook of testing and assessment* (pp. 349–363). Oxford University Press.

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. (2020). *Does multidimensional forced-choice prevent faking? Comparing the susceptibility of the multidimensional forced-choice format and the rating scale format to faking*. PsyArXiv. https://doi.org/10.31234/osf.io/qn5my.

Wetzel, E., & Greiff, S. (2018). The world beyond rating scales – why we should think more carefully about the response format in questionnaires. *European Journal of Psychological Assessment, 34*(1), 1–5. https://doi.org/10.1027/1015-5759/a000469.

Xiao, Y., Liu, H. Y., & Li, H. (2017). **[Author: add page range]** Integration of the forced-choice questionnaire and the Likert scale: A simulation study. *Frontiers in Psychology, 8*. https://doi.org/10.3389/fpsyg.2017.00806.

Zhang, B., Sun, T., Drasgow, F., Chernyshenko, O. S., Nye, C. D., Stark, S., & White, L. A. (2019). Though forced, still valid: Psychometric equivalence of forced-choice and single-statement measures. *Organizational Research Methods, 23*(3), 569–590. https://doi.org/10.1177/1094428119836486.

Ziegler, M. (2014). Stop and state your intentions! Let's not forget the ABC of test construction. *European Journal of Psychological Assessment, 30*(4), 239–242. https://doi.org/10.1027/1015-5759/a000228

Acknowledgement

The authors thank Rebekka Kupffer for her comments on a draft of this editorial.

Eunike Wetzel
Department of Psychology
Otto-von-Guericke University Magdeburg
Universitätsplatz 2
39106 Magdeburg
Germany
eunike.wetzel@ovgu.de