Article

Taking the Test Taker's Perspective: Response Process and Test Motivation in Multidimensional Forced-Choice Versus Rating Scale Instruments

Assessment
2020, Vol. 27(3) 572–584

DOI: 10.1177/1073191118762049
journals.sagepub.com/home/asm
SAGE

Rachelle Sass[1*], Susanne Frick[1], Ulf-Dietrich Reips[1], and Eunike Wetzel[1,2*]

Abstract

The multidimensional forced-choice (MFC) format has been proposed as an alternative to the rating scale (RS) response format. However, it is unclear how changing the response format may affect the response process and test motivation of participants. In Study 1, we investigated the MFC response process using the think-aloud technique. In Study 2, we compared test motivation between the RS format and different versions of the MFC format (presenting 2, 3, 4, and 5 items simultaneously). The response process to MFC item blocks was similar to the RS response process but involved an additional step of weighing the items within a block against each other. The RS and MFC response format groups did not differ in their test motivation. Thus, from the test taker's perspective, the MFC format is somewhat more demanding to respond to, but this does not appear to decrease test motivation.

Keywords

response format, multidimensional forced-choice, rating scale, response process, test motivation, item format, Likert scale

When constructing a self-report questionnaire to assess constructs such as personality traits, interests, or attitudes, test constructors must make many decisions. One important decision pertains to the questionnaire format for collecting item responses. The most commonly applied questionnaire format is a single-stimulus item presentation combined with a rating scale (RS) response format. With an RS response format, respondents rate each item individually on an RS ranging, for example, from *strongly disagree* to *strongly agree* or from *never* to *always* (see example in Figure 1A). The RS format has been criticized widely because of its susceptibility to response biases such as extreme response style, acquiescence, and socially desirable responding (Baumgartner & Steenkamp, 2001; Van Vaerenbergh & Thomas, 2013; Wetzel, Carstensen, & Böhnke, 2013; Wetzel, Lüdtke, Zettler, & Böhnke, 2016). There are some modifications of the RS format such as visual analogue scales (e.g., Funke & Reips, 2012; Kuhlmann, Dantlgraber, & Reips, 2017), but it is unclear whether they have been able to eliminate these biases.[1] Several researchers have proposed using a multidimensional forced-choice (MFC) format as an alternative (Brown & Maydeu-Olivares, 2011; Christiansen, Burns, & Montgomery, 2005). In the MFC format, several items measuring different traits are presented to test takers simultaneously and test takers either rank them according to how well the items describe them or select the one that describes them best and the one that describes them least.[2] For an example, see Figure 1B. The studies described here apply a full ranking format. The MFC format has gained popularity with the development of the Thurstonian item response model (Brown & Maydeu-Olivares, 2011, 2013; Maydeu-Olivares & Brown, 2010), which allows obtaining normative trait estimates as opposed to only ipsative scores as with conventional analysis methods (Brown & Maydeu-Olivares, 2013).

Previous comparisons of the RS and MFC format have focused on aspects of direct relevance to decisions made based on test scores, such as the equivalence of trait estimates and criterion-related validities (Bartram, 2007; Brown & Maydeu-Olivares, 2011, 2013; Heggestad, Morrison, Reeve, & McCloy, 2006). For example, Bartram (2007) compared the validity of forced-choice and RS

[1]University of Konstanz, Konstanz, Germany
[2]Otto-von-Guericke University Magdeburg, Magdeburg, Germany

*Eunike Wetzel and Rachelle Sass are now at the University of Mannheim.

Corresponding Author:

Eunike Wetzel, Department of Psychology, University of Mannheim, L13, 15, 68161 Mannheim, Germany.
Email: eunike.wetzel@uni-mannheim.de

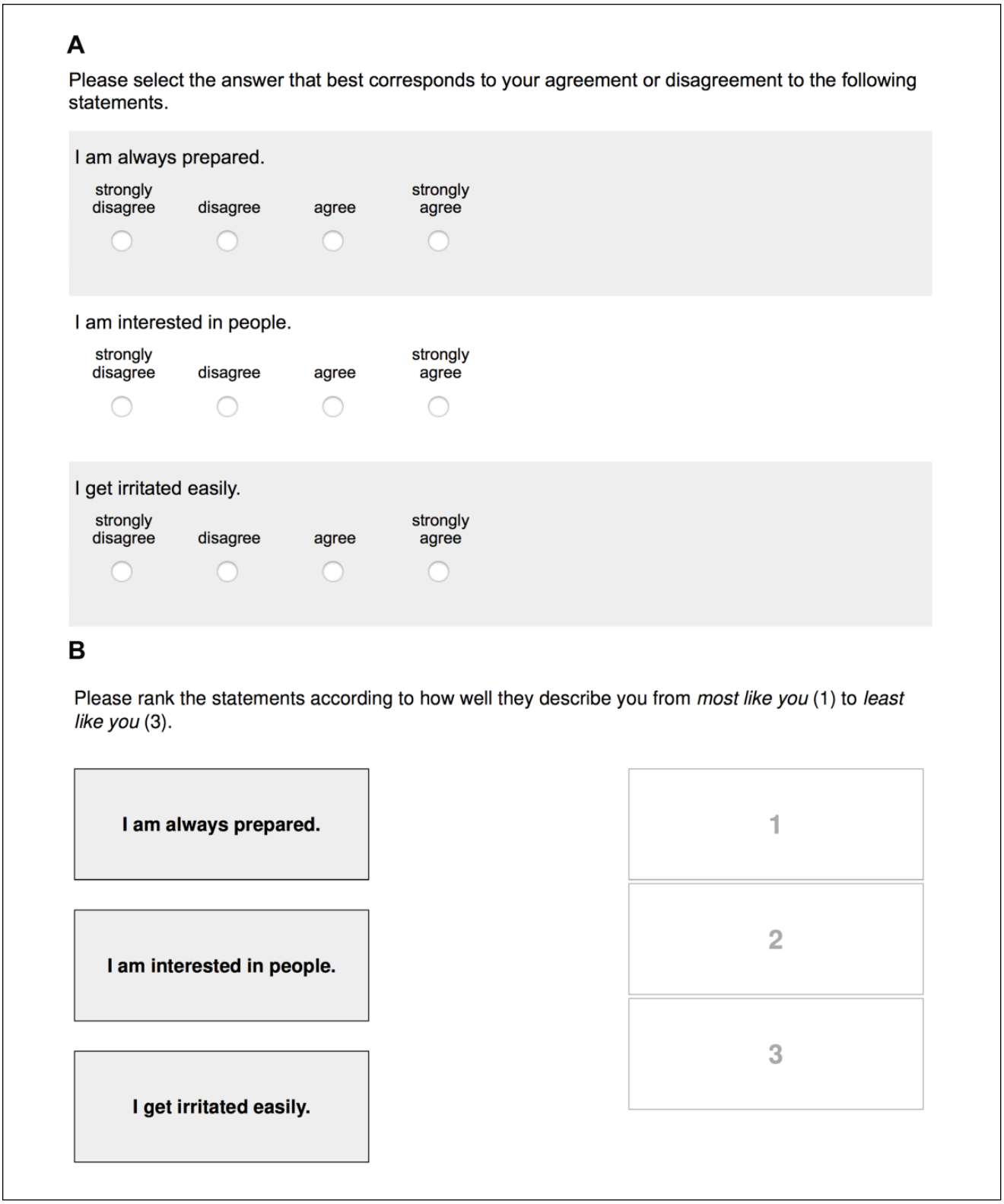

Figure 1. (Panel A) An example for a rating scale format. (Panel B) An example for a multidimensional forced-choice format. In both examples, the first item assesses conscientiousness, the second extraversion, and the third neuroticism.

versions of the Occupational Personality Questionnaire (OPQ; Saville, Holdsworth, Nyfield, Cramp, & Mabey, 1993) for predicting performance ratings across a variety of jobs. He found higher criterion-related validities for the forced-choice version (r = .38) compared with the RS version (r = .25). Salgado and Táuriz (2014) conducted a meta-analysis on the validity of the Big Five assessed with forced-choice questionnaires in predicting occupational criteria and compared these validities with those found in published research with RS questionnaires. They showed that forced-choice questionnaires of the Big Five had similar or higher validities than Big Five questionnaires presented in an RS format. For example, conscientiousness achieved a validity of .24 with forced-choice (Salgado & Táuriz, 2014) and a validity of .23 with RS (Barrick & Mount, 1991) for predicting job performance. For productivity as the criterion, forced-choice showed a slight advantage compared with RS with a validity of .27 versus .17.

One other important consideration concerning the choice of response format should be how test takers respond to it, both in terms of the actual response process underlying the responses to items as well as in terms of their test motivation. If respondents had difficulty using a response format or showed lower test motivation with a response format, decisions based on trait estimates from this response format might be less valid. Thus, this article attempts to take the test taker's perspective to responding to questionnaires. In Study 1, we will compare the RS and MFC format regarding the response process using the think-aloud technique and cognitive interviews. In Study 2, we will compare test-taking motivation between the RS and different versions of the MFC format in an experiment. In the general discussion, we will discuss the implications of the findings from both studies with respect to the application of the RS and MFC formats with self-report questionnaires.

Study 1

There are five stages involved in responding to a questionnaire item presented with the RS format according to Tourangeau and Rasinski (1988) and Podsakoff, MacKenzie, Lee, and Podsakoff (2003): (1) comprehension, (2) retrieval, (3) judgment, (4) response selection, and (5) response reporting.

Each stage requires certain activities. Some of these activities may be deliberate and accessible to conscious awareness, whereas others may be rather automatic and therefore not fully accessible to conscious awareness. In the first stage, respondents perform the most basic steps of attending to the item and instruction. They read the item and *comprehend* the item's content. The *retrieval stage* involves "generating retrieval strategies and cues, retrieving specific and generic memories, and filling in missing details" (Podsakoff et al., 2003, p. 886). During the *judgment stage*, individuals assess how accurate and complete their memories are and draw inferences based on the accessibility of information. The *response selection stage* comprises mapping the judgment onto a particular response category on the RS. Finally, during the *response reporting stage*, respondents actually give the response. At this final stage, adjustments, such as distorting the response in the direction of social desirability, can occur. In the optimal case of high test motivation, respondents will perform all stages carefully and comprehensively. However, with low test motivation, respondents may execute some stages less diligently or even omit them.

Whereas the response process to RS items is well understood, this is not true of the MFC format. We assume that the same general stages are involved in responding to MFC item blocks. However, because the response consists of ranking items, an additional step presumably takes place in the MFC format because of the necessity of weighing the items in the block against each other to determine their rank. The purpose of Study 1 was to gain an understanding of the response process that takes place with the MFC format. To accomplish this, participants were assessed using the think-aloud technique and post hoc cognitive interviews. Because participants were recorded and tested individually in the laboratory, we assume that all of them showed high test motivation and executed all stages of the response process. In the following, we will describe the methods of our study in more detail. Then, based on the results, we will develop a response process stage model for the MFC format. Last, we will discuss in which aspects the response process for the MFC format differs from or is similar to the response process for the RS format.

Method

Participants. The sample consisted of two subsamples. The first subsample comprised 30 students from the University of Konstanz in Germany. Their mean age was 24.6 years (standard deviation [SD] = 5.3 years) and 80% of them were female. Five of the participants completed the study in English and 25 completed the study in German. Psychology students received course credit for participating. The others were remunerated with eight Euros each. The second subsample was composed of 12 adults (11 employed, 1 housewife) from Northeastern Germany. Their mean age was 38.6 years (SD = 11.1 years) and 67% of them were female. All of these participants completed the study in German. They received 15 Euros each for their participation. Thus, the total sample size was 42.

Instruments. We administered 18 MFC triplets assessing the Big Five personality traits (neuroticism, extraversion, openness to experience, agreeableness, conscientiousness) to the participants. These MFC triplets were taken from a pilot

version of the Big Five Triplets instrument (see Study 2). Items within the triplets were matched regarding their social desirability. The instruction for the MFC triplets was "Please rank the statements according to how well they describe you from *most like you* to *least like you*." The three statements were presented on the left side of the computer screen and participants dragged and dropped them into the empty boxes for Ranks 1 to 3 on the right side (see Figure 1B).

Procedure. Participants individually came into the research laboratory.[3] They were instructed to describe their thought processes out loud while they were filling out the questionnaire. After instructing the participants and answering any remaining questions, the research assistant left the room. In order to familiarize participants with the procedure of verbalizing their thoughts as well as to familiarize them with the MFC format, they were given two practice triplets before the actual questionnaire began. After participants had completed the questionnaire and signaled the research assistant that they were done, the research assistant conducted a cognitive interview with the participants and asked them eight open-ended questions to gain further insight into their response process. The participants' voices were recorded throughout the study.

Ratings. The recordings were transcribed by research assistants. Then, the authors read all transcripts and discussed their ideas of what the response process looked like and which elements appeared to be common across participants. We developed a preliminary model of the response process based on this first reading of the transcripts. It was apparent from the transcripts that participants varied in how they responded to the triplets and that parts of the response process were often not verbalized. Thus, we developed a coding scheme to quantify information on the response process of completing the MFC triplets. Specifically, our goal was to quantify how often certain behaviors (e.g., the tendency to form a preliminary judgment of each individual item) occurred as part of the MFC response process within and across participants. With this coding scheme we also wanted to gain insight into the sequence of different stages in the response process. The items in this coding scheme were constructed from the discussion of the transcripts and the preliminary response process model. There were 22 items that the participants were rated on; some sample items include, "Weighs all items against each other before making a judgment" and "Ranks item which describes him/her best first." Refer to Table 1 for a full list of the items in the coding scheme. Participants' behavior with respect to these items was rated on a frequency scale from *never* (1) to *always* (4). Subsample 1 was rated by six raters (five for German-speaking participants) and Subsample 2 was rated by two raters. Interrater reliabilities based on Finn's (1970) coefficient ranged from 0.56 to 0.99 for the 22 items with 86% above 0.70.[4] Final ratings for the analysis were obtained by averaging across raters. In summary, the steps for obtaining the transcript codings were (1) read transcripts, (2) discuss ideas about response process, (3) develop preliminary response process model, (4) develop coding scheme for quantifying information from transcripts, (5) raters code transcripts with coding scheme, and (6) obtain final ratings by averaging across raters. In addition, to quantify information gained from the cognitive interviews, the answers to the eight interview questions were rated by two of the authors using the categories *no, sometimes*, and *yes* (range of interrater reliabilities: .93 to .99). The interview questions can be downloaded at osf.io/k49yt. Finally, the preliminary response process model was revised based on the results of the codings.

Results

Table 1 shows the percentage of participants from the combined subsamples assigned each of the frequency categories from *never* to *always* on the 22 items in the coding scheme. Most participants read each of the three statements in the triplet out loud before verbalizing thoughts related to the decision process. In some cases, participants directly gave a preliminary judgment for each individual statement when they read it out loud as illustrated by the following quote from one participant: "I like routine. That's wrong. I find it easy to get my way, that's wrong. I think a lot when I have to make a decision, that's true. It's more true than I like routines."

Almost all participants (95%) never indicated any difficulties with comprehending the items, although 40% of the participants sometimes expressed uncertainty as to which context the item meaning was referring to. Most participants retrieved relevant information by identifying instances of specific behavior from the past (55% sometimes, 19% usually, and 2% always). This was confirmed by the post hoc cognitive interviews in which 95% of the participants said they retrieved information by identifying instances of specific behavior from the past. Forty-five percent elaborately described instances of specific behavior from the past by going into detail on the situation they were describing. See Table 1 for other behaviors related to retrieval. Thus, the first two stages of the response process in the MFC format can be called—similar to the RS format—*comprehension* and *retrieval*.

The next stage, *judgment*, differs between the response formats in that several statements have to be judged in the MFC format as opposed to just one in the RS format. In addition, there appear to be two distinct pathways in the MFC response process that participants took. These two pathways are illustrated in Figure 2, which shows the response process model for MFC triplets. After comprehending and retrieving information for Item 1 in the triplet,

Table 1. Percentage of Respondents Rated in Each Category per Item.

Item content	*Never* (1)	*Sometimes* (2)	*Usually* (3)	*Always* (4)
Comprehension				
Expresses difficulty with comprehending items	95	5	—	—
Expresses uncertainty as to which context item meaning is referring to	60	40	—	—
Retrieval				
Identifies instances of specific behavior from the past	24	55	19	2
Elaborately describes instances of specific behavior from the past	55	38	5	2
Expresses difficulty when trying to come up with past instances of behavior	88	12	—	—
Identifies instances of behavior which both confirm and disconfirm statements	57	43	—	—
Makes a judgment about an item based on observer accounts of his/her behavior	81	19	—	—
Weighing and ranking of items (judgment)				
Ranks item which describes him/her best first	—	86	12	2
Ranks item which describes him/her least first	2	91	7	—
Ranks item judged as middle first	55	45	—	—
Ranks item which cannot be decided on last	5	95	—	—
Ranks each statement individually, without taking the other statements into account	24	74	2	—
Evaluates single statements individually before comparing final ratings against each other	2	74	24	—
Weighs all items against each other before making judgment	2	43	55	—
Repeats the items out loud several times before making a judgment	9	60	29	2
Judges the items by comparing the second and third items in the triplet to the first item	90	10	—	—
Expresses difficulty comparing all the statements as to which describes him/her most/least	9	81	10	—
Expresses difficulty deciding where to place items because none describe him/her well	38	60	2	—
Expresses difficulty deciding where to place items because all describe him/her well	62	38	—	—
Response selection				
Reconsiders judgment before placing items in ranks	38	62	—	—
Expresses confidence about final judgment	45	55	—	—
Response reporting				
Expresses intent to give ranks that are consistent with ranks given on previous triplets	64	36	—	—

participants sometimes gave a preliminary judgment as in the quote above. If the judgment was clear (agreement or disagreement), this led to the preliminary selection of a rank (1 or 3), either verbally or also by manually moving the item to the respective rank. If the judgment was unclear, the selection of a rank was put on hold and the participant went on to comprehend and retrieve information for the next item (see loop in upper pathway in Figure 2). This procedure was repeated for the remaining two items in the triplet. Then, the items with their preliminary judgments were weighed against each other and the final ranks were selected. On the other hand, sometimes, participants refrained from giving preliminary judgments and instead sequentially comprehended and retrieved information for all items (see lower pathway in Figure 2). Then, they weighed all items against each other before making a judgment and selecting ranks. Thus, the judgment stage in the MFC format appears to sometimes be preceded by and sometimes mixed with *weighing the items* (lower and upper pathway in Figure 2, respectively).

Seventy-six percent of the participants appeared to take the upper pathway including preliminary judgments sometimes or usually. In the interviews, 64% of the participants reported having taken this pathway. Almost all participants (98%) took the lower pathway without preliminary judgments for at least some triplets. Thus, participants did not consistently use one of the two pathways, but rather switched between pathways in the course of the MFC

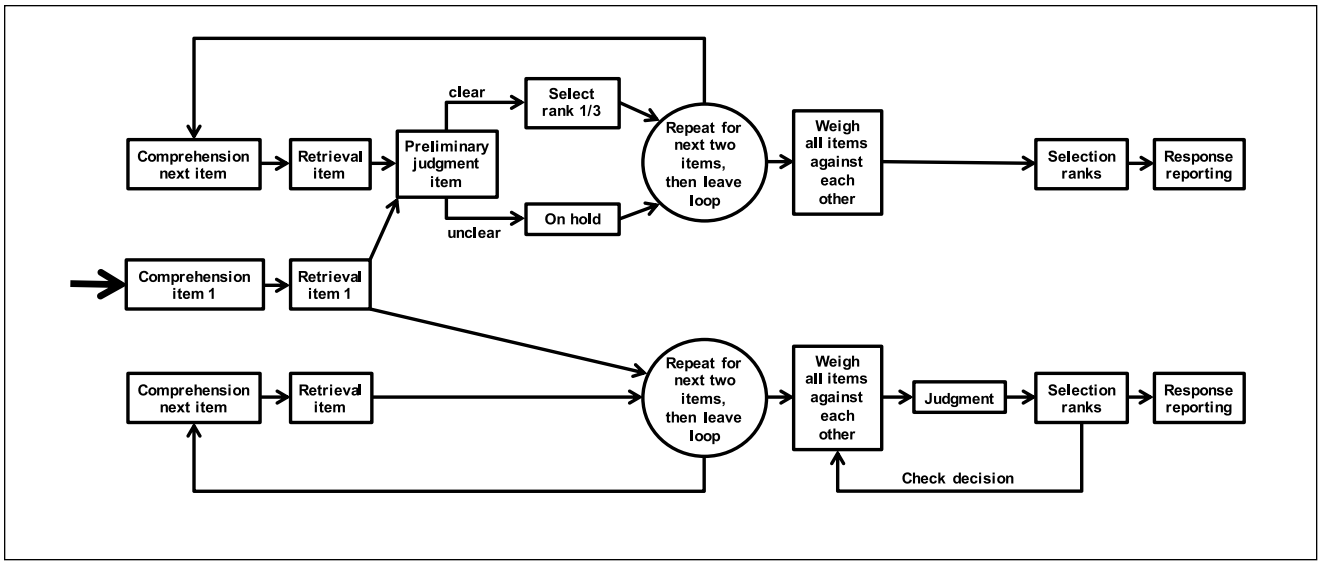

Figure 2. Response process in the multidimensional forced-choice format using triplets. The process begins at the bold arrow on the left. Respondents may take one of two paths depending on whether they judge each item preliminarily after comprehending and retrieving its information (upper pathway) or whether they judge all items together after comprehending and retrieving for all items (lower pathway).

questionnaire. Almost all participants preferred to assign Rank 1 (*most like me*) or Rank 3 (*least like me*) first. Nearly half (45%) of the participants sometimes ranked the item judged as the middle first. Most participants (91%) expressed a general difficulty in comparing all the statements as to which described them best or least for at least some triplets. In particular, 62% of the participants sometimes or usually expressed difficulty deciding where to place the items because none described them well. Furthermore, 38% in some instances expressed difficulty deciding where to place the items because all described them well. For example, one participant said, "So here I would place everything on rank 1, I have to say. Would all describe me really, really well. And I wouldn't make any gradations here, but I did now, because the task requires it." Another participant stated: "So it is really hard to rank this, because I either want to place two things on 1 or two on 3." These ranking difficulties seemed largely not to be caused by high working memory load, because 76% of the participants reported having no difficulty in keeping the retrieved information for the three statements in mind.

Sixty-two percent of the participants sometimes reconsidered their judgment before actually placing the items in ranks (as also indicated by 74% in the interview). This corresponds to the check decision loop in the lower pathway. In the upper pathway, it corresponds to the step during which all items are weighed against each other and the preliminary judgments are reconsidered. Reconsidering sometimes involved pairwise comparisons. For example, one participant said in the interview: "I think that I first compared the top statement with the one in the middle and then figured, ok this fits, and then compared the middle one with the bottom statement to see whether that fits." Fifty-five percent of the participants also expressed confidence about their final judgment (90% indicated that this was the case for some or most triplets in the interview).

The last two stages, *response selection* and *response reporting* appear to be very similar in the MFC format and in the RS format. As is known for the RS format (Podsakoff et al., 2003; Tourangeau & Rasinski, 1998; Tourangeau, Rips, & Rasinski, 2000), response editing also appears to take place at the response reporting stage in the MFC format: Thirty-six percent of the participants sometimes expressed their intent to give ranks that were consistent with ranks they had given on previous triplets. For example, one participant said "I also love big parties, but because of the previous task it would not be transitive or somewhat illogical if I placed I love big parties on one. That is the reason I'd place it on three."[5]

Discussion

Overall, as we expected, the response process in the MFC format appeared to be quite similar to the response process in the RS format. However, importantly, the MFC format involved an additional stage of weighing the items against each other. This stage appeared to come after all items in the triplet had been comprehended and relevant information retrieved. In one pathway of the MFC response process model, the weighing phase could be clearly distinguished

from the subsequent judgment phase. However, in the other pathway, it could not be clearly distinguished from the judgment phase because often items were given a preliminary judgment, which was later reconsidered when all items were taken into account. One factor that may influence the weighing and judgment phase is how easily a respondent can rank the items with respect to how well they describe him or her (i.e., the distance between items). With smaller distances between items, more weighing and deliberation may need to take place. This is a further difference to the RS format where each item is considered individually. Note, however, that in the RS format respondents will also sometimes consider their responses to previous items in determining their response to a new item (*consistency effects*).

Of course, the think-aloud technique has its limitations, including that the act of verbalizing one's thoughts may change the response process taking place and that people may differ in their ability to verbalize. In addition, it is conceivable that not all aspects of the response process can be verbalized equally well. Some parts of the response process, in particular activities related to retrieval and judgment, may not even be accessible to conscious awareness. As suggested by Tourangeau and Rasinski (1988), whether processes take place in a controlled, explicit manner or automatically may depend on the accessibility of the content. For example, reconsider the quote from one of our participants from above: "I like routine. That's wrong. I find it easy to get my way, that's wrong. I think a lot when I have to make a decision, that's true. It's more true than I like routines." To make the preliminary judgment that "I like routine" is "wrong," some retrieval must have taken place, but this was not verbalized. Thus, our MFC response process model probably represents a strong simplification of the cognitive processes underlying responses to MFC triplets. In addition, while the temporal sequence suggested by our model is plausible based on respondents' comments and interview responses and is in line with previous research on the RS format, we cannot be certain that the steps take place in this order. Again, the model is a simplification and it is quite likely that the true processes underlying MFC responses are a lot more complex and may include switching around between steps. Nevertheless, the model is useful because it allows us to gain a first insight into the MFC response process, which will hopefully prompt further research.

This study investigated a particular type of MFC format, namely one using triplets. One open question is whether the response process is the same when other versions of the MFC format are applied such as pairwise comparisons or quads. Furthermore, regarding the two pathways in our model, it is unclear which factors influence which pathway is taken and whether there are individual differences in the preference for a certain pathway. The preliminary judgment given to individual items in one pathway appears to be similar to a rating of the items with a dichotomous true/false response format. It also appears to be similar to the agree/disagree decision that is part of item response tree models of the process of responding to RS items (Böckenholt, 2012). In the Thurstonian item response model for analyzing MFC data, ranks are transformed into binary outcome variables based on pairwise comparisons between all items in the triplet (see Brown & Maydeu-Olivares, 2011). Based on our results, this modeling method appears justified because participants weigh the items against each other in deciding their ranks and often explicitly compare two items at a time.

In sum, the response process to MFC triplets appears to require the same general stages as the response process to items presented in the RS format, though more research on the topic is needed. An additional stage takes place with the MFC format, however, which is the weighing of the items against each other. As noted above, many participants expressed difficulty in comparing the statements and deciding how to rank them. Thus, another important aspect when considering the test taker's perspective is in how far this may affect test motivation, which is what we turn to next.

Study 2

The goal of Study 2 was to investigate whether test takers' test motivation differs between the RS and different versions of the MFC format. *Test motivation* can be defined as "the willingness to engage in working on test items and to invest effort and persistence in this undertaking" (Baumert & Demmrich, 2001, p. 441). There are a variety of factors that may affect test motivation, most importantly, task difficulty and cognitive load. In particular, if task difficulty and cognitive load are high, test motivation will be low (Krosnick, 1991). Another factor that can affect test motivation is the face validity of the instrument. When participants perceive the content of test items as being relevant and meaningful, they may be more fully engaged (Krosnick, 1991). In addition, whether the assessment context is low-stakes or high-stakes can play a role. For example, Wolf and Smith (1995) found that when college students completed two forms of a course exam (one of which counted toward their class grade and one of which did not), the condition in which the exam counted toward their grade (high-stakes) was shown to be associated with significantly higher performance as well as significantly higher test motivation. Nevertheless, test motivation can also be high in low-stakes contexts as shown by Eklöf (2007) for Swedish students participating in the Trends in International Mathematics and Science Study (TIMSS) and by Baumert and Demmrich (2001) for German students taking part of the Programme for International Student Assessment (PISA) mathematical literacy test, both of which are low-stakes situations for the individual students. Furthermore, incentives appear to

increase test motivation (Baumert & Demmrich, 2001; Duckworth, Quinn, Lynam, Loeber, & Stouthamer-Loeber, 2011; Marsh, 1984). For example, Duckworth et al. (2011) showed that incentives in random-assignment studies were related to higher IQ scores.

Personality traits also appear to influence test motivation. Salgado, Remeseiro, and Iglesias (1996) differentiated two independent facets of test motivation: one characterized by a negative belief about tests (e.g., high anxiety in test situations) and the second characterized by a positive belief that an individual perceives the test as a fair, motivating, and attractive procedure for making personnel decisions. They found that neuroticism and agreeableness were positively related to the negative facet of test-taking motivation. On the other hand, extraversion, openness to experience, and conscientiousness were found to be positively related to the positive facet of test-taking motivation. Last, test motivation can also have a moderating effect on the predictive validity of IQ scores and personality scores (Duckworth et al., 2011; O'Neill, Goffin, & Gellatly, 2010; Schmit & Ryan, 1997). For example, Duckworth et al. (2011) found that observer ratings of test motivation were associated with both IQ scores and important life outcomes; children who tried harder on the low-stakes test received higher IQ scores and also had more positive life outcomes. Thus, test motivation is an important influence on test scores in psychological assessment.

As the results from Study 1 showed, the response process appears to be quite similar between the RS and MFC format except that the MFC format additionally requires weighing the items in a block against each other. As the items are weighed against each other, relevant information needs to be kept active in working memory. The MFC triplets, composed of three items, did not pose a problem for most participants as shown in Study 1. However, the more items are presented simultaneously and need to be weighed against each other, the higher the demands on working memory should be. Based on previous research (e.g., Krosnick, 1991), this increasing demand may result in declining test motivation. Moreover, test motivation may also be lower in the MFC format because respondents are forced to make a decision between items and cannot place two items on the same rank whereas they can respond with the same category to both items in the RS format. Some participants in Study 1 expressed frustration with having to rank items (see above). Thus, in Study 2, we will compare test motivation not only between the RS format and the MFC triplets format, but we will additionally vary the number of items presented simultaneously in the MFC format. In particular, blocks of two (pairs), three (triplets), four (quads), and five (pentads) will be applied. Based on the results of Study 1, we do not expect differences in test motivation between RS and MFC pairs and MFC triplets. However, considering the additional complexity and cognitive effort involved in ranking four or five items, we hypothesize that the conditions of MFC quads and MFC pentads will show lower test motivation than the RS format, where only one item needs to be evaluated at a time.

Method

Participants. Data were collected online and participants were recruited by posting the study link on various web pages, including Facebook pages, the department's website, and several other websites. Individuals who completed the questionnaire online (1,044 participants) had the opportunity to take part in a lottery of 20 Amazon gift cards worth 25, 50, 75, and 100 Euros. According to our a priori power analysis, approximately 2,000 participants (400 per group) were necessary to achieve a power of .90 assuming a small to moderate effect size. To recruit the remaining participants, we also collected data via two Internet access panels: *Prolific Academic* and *Respondi*. Prolific Academic is a U.K.-based platform similar to Amazon Mechanical Turk. That is, participants sign up and choose studies to participate in. Prolific Academic participants appear to be more naïve and more diverse than MTurk participants (Peer, Brandimarte, Samat, & Acquisti, 2017). Respondi participants once register to the panel and are then invited to selected studies via e-mail. The quality of their data is checked regularly. In both cases, participants receive individual payments for taking part in studies. Participants recruited through Prolific Academic (n = 191) were compensated with 1.10 British pounds each for their participation. Respondi participants (n = 1,156) were remunerated with 1.00 Euro each.

Thus, in total, we collected data on 2,391 participants. Of these, data from 224 participants were excluded because of several criteria: Data from 69 participants were excluded because they reported having already taken part in the study (the default option on the first item in the questionnaire). Of the remaining participants, 29 were excluded because they reported having taken part in a similar study at the end of the questionnaire (presumably a different study applying the same instrument or a different response format version in the same study). Data from 32 participants were excluded for filling out the questionnaire too quickly (more than 2 *SD* below the average). Last, data from 94 participants who failed an instructed response item were excluded. The final sample size therefore consisted of 2,167 participants. Of those, 52% were female with a mean age of 37 years (SD = 15). Table 2 displays participant demographic characteristics by response format group. The data from the final sample are available on the Open Science Framework (osf.io/kqady).

Measures. We applied the Big Five Triplets (Wetzel & Frick, 2017), which assess the Big Five personality traits

Table 2. Participant Demographic Characteristics and Test Motivation Mean Scores in Study 2.

	N	M_{age} (*SD*)	Percent female	$M_{test\ motivation}$ (*SD*)
Total	2,167	37 (15)	52	3.35 (0.39)
By condition				
Rating scale	409	36 (14)	50	3.30 (0.39)
Pairs	454	37 (15)	55	3.34 (0.39)
Triplets	463	38 (16)	52	3.37 (0.39)
Quads	425	36 (15)	53	3.36 (0.40)
Pentads	416	37 (15)	50	3.36 (0.40)

neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness. The original response format is an MFC format with 20 triplets (i.e., 60 items in total). This is the same format as the one applied in Study 1 (see Figure 1B for an example). For the purposes of this study, we varied the format of the Big Five Triplets to construct the different response format versions. The number of individual items (60) presented in each version was held constant. For the RS version, items were presented individually (three on one page). For the alternations of the MFC format, items were allocated to 30 pairs, 15 quads, and 12 pentads, respectively.

Test motivation was assessed using items from the Test Attitude Survey (Arvey, Strickland, Drauden, & Martin, 1990) adapted to the context of this study, as well as some additional items constructed by the authors, totaling 19 items.[6] The items measured participants' motivation, concentration, and enjoyment on a four-point RS ranging from *strongly disagree* to *strongly agree*. A sample item from the motivation subscale is, "I tried to be as accurate as I could be on this questionnaire"; a sample item from the concentration subscale is, "I get distracted when taking surveys of this type"; and a sample item from the enjoyment subscale is, "I greatly enjoyed filling out this questionnaire." For a complete list of the items on the Test Motivation Questionnaire, see the supplemental material available in the online version of the article. McDonald's omega (McDonald, 1999) for the test motivation scores was .93. Participants in one of the MFC groups also filled out four items assessing their ease of ranking the items such as "I found it easy to rank the statements in each block according to how well they described me." All participants additionally filled out a few demographic questions. All samples except the Respondi sample also filled out the Short Dark Triad questionnaire (Jones & Paulhus, 2014) for the purposes of a different study.

Procedure. Participants were randomly assigned to one of the five questionnaire formats: RS, MFC pairs, MFC triplets, MFC quads, and MFC pentads (see Table 2 for a breakdown of sample size by condition). Respondents first completed the Big Five Triplets in one of the questionnaire versions and directly afterward the Test Motivation Questionnaire.

Analysis. We first explored whether dropout rates differed between response format groups. To address our research question of whether differences in test motivation existed between response format groups, we computed an analysis of variance in R (R Core Team, 2013) with the test motivation mean score as the dependent variable.

Results

Dropout rates did not differ between response format groups ($\chi^2 = 5.13$, $p = .27$). As the mean scores by group in Table 2 show, no differences in test motivation were found between the response format groups, $F(4, 1974) = 2.23$, $p = .06$. All pairwise mean differences between the RS group and an MFC group yielded Cohen's *d* values close to 0. Thus, our hypothesis that participants taking the MFC quads or MFC pentads version of the questionnaire would report lower test motivation than participants taking the RS version was not confirmed. However, our expectation that the RS group would not differ from the MFC pairs group and the MFC triplets group regarding their test motivation scores was confirmed. Results did not differ between Internet access panel participants and those from other Internet sources. Furthermore, results were robust when we included participants who had not passed the instructed response item.

Discussion

All response format groups reported equally high test motivation. This finding is particularly interesting because in Study 1, many participants expressed difficulty with ranking items and some also expressed exasperation and frustration with being forced to select ranks. In addition, in Study 2, 40% of the participants indicated disagreement with the statement "I found it easy to rank the statements in each block according to how well they described me." Possible reasons for the lack of differences in test motivation between response format groups include differences in the length of the questionnaire. While the number of items was constant

across response format versions, the number of pages in the questionnaire differed: For both the RS and the MFC triplets versions, the Big Five questionnaire consisted of 20 pages (three individual items or one triplet per page, respectively). In contrast, with MFC quads, the number of pages equaled 15 and with MFC pentads, it equaled 12. In this context, it is important to consider one possible indicator for the cognitive load involved with responding to the items, namely the average time taken to complete the questionnaire. A comparison of the average time participants took to complete the questionnaire showed that it was lowest in the RS group ($M = 8.3$ minutes, $SD = 3.0$) and highest in the MFC quads and pentads groups ($M = 9.5$ minutes, $SD = 2.8$ and $M = 9.3$ minutes, $SD = 3.6$, respectively). A contrast of the RS group against the average of the MFC quads and pentads groups indicated that this difference was significant, $t(1347) = -4.97$, $p < .001$. Thus, despite the shorter questionnaire length, participants in the MFC quads and MFC pentads groups took slightly longer to fill out the questionnaire than participants in one of the other groups, in particular the RS group, indicating that cognitive load may have been higher the more statements were presented simultaneously in a block. It is conceivable that the shorter quads and pentads questionnaire may have cancelled out the negative effect of higher cognitive load on test motivation.

Another reason why the MFC format with quads and pentads may not have negatively affected test motivation despite the higher cognitive load may be the novelty effect of the MFC format. Presumably, for most participants this was the first time they filled out a questionnaire in the MFC format, and this may have increased their interest and test motivation. However, we did not ask participants whether they had come across the MFC format before, so this is speculation.

One limitation of this study is that we recruited participants with different methods, namely participants from Internet access panels and participants from other Internet sources. These different sample types had different incentives: Access panel participants were individually remunerated, whereas participants from other Internet sources had the chance of winning a gift card. Nevertheless, participants from the two sample types did not differ in their average test motivation scores. Another limitation is that self-reported motivation may not accurately reflect actual motivation. This is a general limitation of the self-report method. We tried to encourage honest responses to the test motivation items by ensuring participants that their responses to these items would not affect their remuneration or their chances of winning a voucher.

In sum, we did not find a difference in test motivation between the RS format and different versions of the MFC format (pairs, triplets, quads, pentads). Switching from the RS format to an MFC format therefore does not appear to come at the cost of lower test motivation.

General Discussion

From the test taker's perspective, completing an MFC questionnaire appears to be slightly more demanding than filling out an RS questionnaire. The response process involves weighing the items against each other before selecting their ranks. While items are weighed against each other, relevant information for all items needs to be kept active in working memory. This cognitive demand was not too challenging with three items in a block as in Study 1: 74% of the participants reported having no difficulty in keeping the retrieved information for three statements in mind. However, it may be more taxing the more items are presented in a block, as indicated by higher completion times with MFC quads and MFC pentads compared with the RS group in Study 2. The tendency of participants to sometimes give preliminary judgments for items separately before taking all of them into consideration together might be a strategy to deal with the higher cognitive demands of the MFC format. Visualizing preliminary judgments by moving items to ranks might be another strategy. Furthermore, as the response format's name suggests, participants are forced to make a choice. They have to rank the items and this—according to participant statements—can be challenging and also somewhat frustrating.

Future research could investigate ways of making the MFC format more test taker friendly. For example, with quads or pentads other instructions could be used instead of requiring a full ranking as in Study 2, such as selecting the item that is most and least like the respondent (see e.g., Brown & Maydeu-Olivares, 2013). This would presumably decrease cognitive load although it comes at the cost of reduced precision in estimating model parameters (Brown & Maydeu-Olivares, 2011). In addition, relations between test motivation and individual differences variables such as personality traits (see e.g., Salgado et al., 1996) and working memory capacity could be investigated. It would also be interesting to investigate the two pathways in the MFC response process in more detail, for example, by applying mouse-tracking and other paradata in Internet-based research (e.g., Stieger & Reips, 2010), and to investigate the response process in other (e.g., high-stakes) assessment contexts. A limitation of the current investigation is that all participants were adults. Thus, future research could investigate the MFC response process and test motivation in children. Note, however, that in particular with young children, self-reports are generally problematic (Soto, John, Gosling, & Potter, 2008). In addition, it would also be interesting to investigate the response process and test motivation in samples from other cultures.

The MFC format eliminates biases due to individual differences in using the RS, including individual differences in interpreting numerical or verbal anchors as well as response

styles such as extreme response style or acquiescence response style. However, other response biases might still occur. For example, at the response reporting stage, participants might rank statements in a way that is consistent with previous triplets or they might rank statements according to their perceived social desirability. Previous research suggests that the MFC format may be less susceptible to the impression management component of socially desirable responding (Paulhus, 2002) than the RS format when comparing groups with a fake good versus a neutral instruction (Christiansen et al., 2005; Heggestad et al., 2006; Jackson, Wroblewski, & Ashton, 2000). However, at the individual level, the same amount of faking may occur in both response formats (Heggestad et al., 2006). It is unclear whether there are any differences between the response formats in their susceptibility to socially desirable responding occurring outside conscious awareness (the self-deception component; Paulhus, 2002). Furthermore, other self-report biases such as careless responding might also occur in the MFC format. It is unclear whether there are differences between the MFC and RS format in the occurrence of these response biases. Recent research indicates that the MFC format may be an adequate method of controlling for some other-report biases common to the RS format such as halo effects (Brown, Inceoglu, & Lin, 2017). Nevertheless, more research comparing the occurrence of response biases that can occur in both formats is needed, for example in terms of frequency. Future research on the MFC format could also investigate potential response biases that might be particular to this format.

Overall, the complexities and demands of responding to MFC items do not appear to impede test motivation. All response format groups we investigated reported equally high test motivation levels. This indicates that switching from an RS format to an MFC format might not come at the cost of lower test motivation, which is an important concern, in particular, in low-stakes testing assessment contexts such as online self-report questionnaires. Instead, it might be a promising strategy to avoid problems inherent to the RS format such as response styles.

Acknowledgments

We thank Theresa Falter, Veronika Held, Clara Jupe, and Felix Lang for their help with data collection and transcribing participant recordings in Study 1.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by a grant from the German Research Foundation (DFG) to Eunike Wetzel (WE 5586/2-1) as well as by the Young Scholar Fund of the University of Konstanz.

Notes

1. The authors disagree on whether visual analogue scales are an improvement over RS.
2. The MFC format is both an item format and a response format. For simplicity in comparing it with RSs, we refer to it as a response format in the following.
3. Participants in Subsample 2 were tested in their homes. The rest of the procedure was the same as with Subsample 1.
4. Finn's (1970) coefficient was chosen instead of the more common intraclass correlation because it corrects for low variances, which was the case for several items.
5. When we analyzed the two subsamples separately, the results overall were very similar. One noteworthy difference is that Subsample 1 (mainly students) overall verbalized more than Subsample 2 (mainly employed adults) and tended to describe more specific situations or examples of behaviors that they thought of when ranking the items.
6. We conducted two pilot studies to validate our set of test motivation items. In the first study, we used the true/false format as in Arvey et al. (1990). Several items showed ceiling effects with more than 90% of the participants indicating "true." Because of these ceiling effects and the low variance, the factor structure was unclear. For the second pilot study, we added some more difficult items such as "This questionnaire was so interesting that I would have liked to continue with it once I arrived at the end." Furthermore, we changed the response format to a four-point scale from *strongly disagree* to *strongly agree*. With these changes fewer items showed ceiling effects and the factor analysis yielded three facets, which we interpreted as motivation, concentration, and enjoyment. All items also showed standardized loadings greater than .40 on an overall test motivation factor in a one-factor model.

Supplemental Material

Supplementary material for this article is available at http://journals.sagepub.com/doi/suppl/10.1177/1073191118762049

References

Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology*, *43*, 695-716.

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, *44*, 1-26.

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

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education*, *16*, 441-462.

Baumgartner, H., & Steenkamp, J.-B. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, *28*, 143-156.

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods*, *17*, 665-678. doi:10.1037/a0028111

Brown, A., Inceoglu, I., & Lin, Y. (2017). Preventing rater biases in 360-degree feedback by forcing choice. *Organizational Research Methods*, *20*, 121-148. doi:10.1177/1094428116668036

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

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

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

Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R., & Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 7716-7720. doi:10.1073/pnas.1018601108

Eklöf, H. (2007). Test-taking motivation and mathematics performance in TIMSS 2003. *International Journal of Testing*, *7*, 311-326.

Finn, R. H. (1970). A note on estimating reliability of categorical data. *Educational and Psychological Measurement*, *30*, 71-76. doi:10.1177/001316447003000106

Funke, F., & Reips, U.-D. (2012). Why semantic differentials in Web-based research should be made from visual analogue scales and not from 5-point scales. *Field Methods*, *24*, 310-327. doi:10.1177/1525822X12444061

Heggestad, E. D., Morrison, M., Reeve, C. L., & McCloy, R. A. (2006). Forced-choice assessments of personality for selection: Evaluating issues of normative assessment and faking resistance. *Journal of Applied Psychology*, *91*, 9-24. doi:10.1037/0021-9010.91.1.9

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*, 371-388. doi:10.1207/S15327043HUP1304_3

Jones, D. N., & Paulhus, D. L. (2014). Introducing the short Dark Triad (SD3): A brief measure of dark personality traits. *Assessment*, *21*, 28-41. doi:10.1177/1073191113514105

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*, 213-236. doi:10.1002/acp.2350050305

Kuhlmann, T., Dantlgraber, M., & Reips, U.-D. (2017). Investigating measurement equivalence of visual analogue scales and Likert-type scales in Internet-based personality questionnaires. *Behavior Research Methods*. Advance online publication. doi:10.3758/s13428-016-0850-x

Marsh, H. W. (1984). Experimental manipulations of university student motivation and their effects on examination performance. *British Journal of Educational Psychology*, *54*, 206-213.

Maydeu-Olivares, A., & Brown, A. (2010). Item response modeling of paired comparison and ranking data. *Multivariate Behavioral Research*, *45*, 935-974. doi:10.1080/00273171.2010.531231

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum.

O'Neill, T. A., Goffin, R. D., & Gellatly, I. R. (2010). Test-taking motivation and personality test validity. *Journal of Personnel Psychology*, *9*, 117-125. doi:10.1027/1866-5888/a000012

Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. I. Braun, D. N. Jackson & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 49-69). Mahwah, NJ: Lawrence Erlbaum.

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153-163. doi:10.1016/j.jesp.2017.01.006

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*, 879-903. doi:10.1037/0021-9101.88.5.879

R Core Team. (2013). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org/

Salgado, J. F., Remeseiro, C., & Iglesias, M. (1996). Personality and test taking motivation. *Psicothema*, *8*, 553-562.

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*, 3-30. doi:10.1080/1359432X.2012.716198

Saville, P., Holdsworth, R., Nyfield, G., Cramp, L., & Mabey, W. (1993). *Occupational Personality Questionnaires: Concept model manual and user's guide*. Esher, England: Saville & Holdsworth Ltd.

Schmit, M. J., & Ryan, A. M. (1997). Applicant withdrawal: The role of test-taking attitudes and racial differences. *Personnel Psychology*, *50*, 855-876. doi:10.1111/j.1744-6570.1997.tb01485.x

Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2008). The developmental psychometrics of big five self-reports: acquiescence, factor structure, coherence, and differentiation from ages 10 to 20. *Journal of Personality and Social Psychology*, *94*, 718-737. doi:10.1037/0022-3514.94.4.718

Stieger, S., & Reips, U.-D. (2010). What are participants doing while filling in an online questionnaire: A paradata collection tool and an empirical study. *Computers in Human Behavior*, *26*, 1488-1495. doi:10.1016/j.chb.2010.05.013

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin*, *103*, 299-314. doi:10.1037/0033-2909.103.3.299

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge, England: Cambridge University Press.

Van Vaerenbergh, Y., & Thomas, T. D. (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research*, *25*, 195-217. doi:10.1093/Ijpor/Eds021

Wetzel, E., Carstensen, C. H., & Böhnke, J. R. (2013). Consistency of extreme response style and non-extreme response style across traits. *Journal of Research in Personality*, *47*, 178-189. doi:10.1016/j.jrp.2012.10.010

Wetzel, E., & Frick, S. (2017). *The Big Five Triplets—Development of a multidimensional forced-choice questionnaire*. Manuscript in preparation.

Wetzel, E., Lüdtke, O., Zettler, I., & Böhnke, J. R. (2016). The stability of extreme response style and acquiescence over 8 years. *Assessment*, *23*, 279-291. doi:10.1177/1073191115583714

Wolf, L. F., & Smith, J. F. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education*, *8*, 227-242. doi:10.1207/s15324818ame0803_3