When middle really means 'top' or 'bottom'

An analysis the 16PF5 using Bock’s nominal response model

Murray, A.L.; Booth, T.; Molenaar, D.

DOI
10.1080/00223891.2015.1095197

Publication date
2016

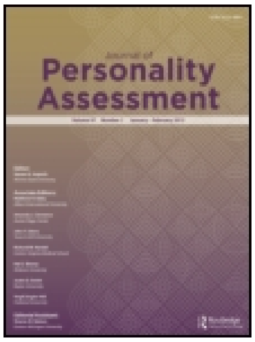
Document Version
Final published version

Published in
Journal of Personality Assessment

License
Article 25fa Dutch Copyright Act

Link to publication

Citation for published version (APA):
Murray, A. L., Booth, T., & Molenaar, D. (2016). When middle really means 'top' or 'bottom': An analysis the 16PF5 using Bock’s nominal response model. *Journal of Personality Assessment*, *98*(3), 319-331. https://doi.org/10.1080/00223891.2015.1095197



Routledge
Taylor & Francis Group

Journal of Personality Assessment

ISSN: 0022-3891 (Print) 1532-7752 (Online) Journal homepage: http://www.tandfonline.com/loi/hjpa20

When Middle Really Means “Top” or “Bottom”: An Analysis of the 16PF5 Using Bock's Nominal Response Model

Aja Louise Murray, Tom Booth & Dylan Molenaar

To cite this article: Aja Louise Murray, Tom Booth & Dylan Molenaar (2016) When Middle Really Means “Top” or “Bottom”: An Analysis of the 16PF5 Using Bock's Nominal Response Model, Journal of Personality Assessment, 98:3, 319-331, DOI: 10.1080/00223891.2015.1095197

To link to this article: http://dx.doi.org/10.1080/00223891.2015.1095197

Published online: 07 Nov 2015.

JOURNAL OF PERSONALITY ASSESSMENT
2016, VOL. 98, NO. 3, 319–331
http://dx.doi.org/10.1080/00223891.2015.1095197

Routledge
Taylor & Francis Group

When Middle Really Means “Top” or “Bottom”: An Analysis of the 16PF5 Using Bock’s Nominal Response Model

Aja Louise Murray,[1,2] Tom Booth,[2] and Dylan Molenaar[3]

[1]Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, United Kingdom; [2]Department of Psychology, University of Edinburgh, United Kingdom; [3]Department of Psychology, University of Amsterdam, The Netherlands

ABSTRACT

When self-report items with a Likert-type scale include a middle response option (e.g., *Unsure, Neither agree nor disagree*, or *?*), this middle option is assumed to measure a level of the trait intermediate between the high and low response categories. In this study, we tested this assumption in the 16 Personality Factor Questionnaire, Version 5 (16PF5) by fitting Bock’s nominal response model in the U.S. and UK standardization samples of the 16PF5. We found that in many cases, the middle option was indicative of higher levels of the latent trait than the ostensibly highest response option. In certain other cases, it was indicative of lower levels of the latent trait than the ostensibly lowest response option. This undermines the use of a simple successive integer scoring scheme where responses in adjacent response categories are assigned scores of 0, 1, and 2. Recommendations for alternative scoring schemes are provided. Results also suggested that certain personality traits, especially neurotic traits, are associated with a tendency toward selecting the middle option.

ARTICLE HISTORY
Received 16 October 2014
Revised 31 July 2015

In inventories measuring psychological constructs, it is common to offer respondents a middle option on the response scale. These middle options have verbal labels such as *not sure, neutral, ?*, or *neither agree nor disagree*, and are typically treated as if a response to this category indicates an intermediate level of the construct, or latent trait, that the inventory purports to measure (Kulas, Stachowski, & Haynes, 2008). For example, if an item has three options, *disagree, ?*, and *agree*, then the three responses are usually assigned scores of 0, 1, and 2, respectively. However, the assumption that middle response options represent an intermediate position on the latent trait continuum is rarely, if ever, assessed and the consequences of its violation are potentially important (González-Romá & Espejo, 2003; Preston, Reise, Cai, & Hays, 2011). In this study, we explored this issue using the 16 Personality Factor Questionnaire, Version 5 (16PF5; Conn & Rieke, 1994).

Research on the use of middle response categories suggests that ambiguous response options such as *?* might be problematic because respondents’ understanding and use of this category does not tend to be consistent (Hernández, Drasgow, & González-Romá, 2004; Kulas & Stachowski, 2013). Rather, it ends up behaving as a “catch-all” response option that is selected when, for whatever reason, the other response options are not viewed as appropriate descriptions of the self. That is, instead of representing an intermediate level of the latent trait as intended, selecting the middle option could represent a multitude of other factors such as ambivalence, indifference, a lack of understanding of the question, a reticence about expressing a trait (e.g., when the items pertain to sensitive topics or socially undesirable traits), or a feeling of being insufficiently informed or familiar to answer the question definitively (DuBois & Burns, 1975; Gonzáles-Romá & Espejo, 2003). Consistent with this, studies have shown that selecting the middle option is associated with scoring high on impression management (Hernández et al. 2004), poorer item clarity (Kulas & Stachowski, 2009), and will often be selected in cases where the true response might be closer to “not applicable” or “it depends” (Kulas et al. 2008; Kulas & Stachowski, 2013). Kulas et al. (2008), for example, found that when individuals were offered two versions of a personality inventory, one with a response option labeled N/A and one without, those individuals who selected the N/A option were likely to select the middle response option in the second version in lieu of an N/A option. These studies suggest that selecting the middle option can reflect a diversity of factors other than being intermediate in the latent trait.

The 16PF5 assesses 15 primary personality scales with a range of 10 to 14 items per scale. Each of these items has a three-point response format that includes a positive and negative option for each trait, and a middle *?* option. Items are scored 0 (*negative option*), 1 (*?*), and 2 (*positive option*). In previous editions of the inventory, the middle option varied across items. However, as noted by Conn and Rieke (1994) in the test manual, this led to complaints by respondents as, for example, they objected to using a middle option of *Uncertain* when they were certain what their response was, but that response did not correspond to either of the other options available to them (p. 8). Therefore, the *?* option was introduced as a way of resolving the ambiguities caused by varied middle response options. In line with the research noted earlier, Conn and Rieke

CONTACT Tom Booth ✉ tom.booth@ed.ac.uk Department of Psychology, University of Edinburgh, 7 George Square, Edinburgh, EH8 9JZ, UK.

acknowledged the possibility of varied use of the *?* option in the 16PF5, stating that this option provides “a uniform response choice that can cover several different reasons for not selecting either the *a* or *c* alternative” (p. 8), where *a* and *c* represent the positive (2) and negative (0) response options for each item.

Selecting the *?* response option for reasons other than being intermediate on the latent trait undermines the functioning of a simple scoring scheme that assigns successive integers to the low (0), middle (1), and high (2) response options. This is because when the sum or average of all items in a scale is used as a score to represent an individual’s position on a latent trait, an implicit scoring scheme is employed wherein each response category is given equal weight and contributes to the sum score in accordance with the numerical value that it is assigned. This assumes that not only are the response categories in the order that corresponds with their assigned numerical values, but that each response category carries equal information about an individual’s position on the latent trait. If either assumption is false—possibly due to one of the response processes described earlier—the sum score will be a distorted representation of an individual’s standing on the latent trait.

The use of the middle response option in an unintended manner is most problematic for sum or average scoring when the selection of the middle response option is related to possessing high or low levels of the very latent trait that is being measured by the item. As an example, this could occur if highly neurotic individuals are more likely to be uncertain about their response and thus tend to disproportionately select the middle option. In this and similar cases, the middle response option might behave as though it is positioned below an ostensibly lower response option or above an ostensibly higher response option on the latent trait continuum. Thus, a successive integer scoring scheme that treats the middle option as lying at an intermediate level of the latent trait will assign the integers in an order that is at odds with their true order and lead to systematic under- or over estimates or trait levels for some individuals.

Despite the potential complications entailed by the *?* option, empirical studies using the 16PF5 tend to employ either this simple successive integer scoring strategy or a similar strategy that makes untested assumptions about the functioning of the *?* response option. Reviewing a number of studies using the 16PF5, it can be seen that most cases do not report their treatment of the *?* option even though the sum scores are used to represent the latent traits measured by the scale (e.g., Aluja, Blanch, & García, 2005; Dancer & Woods, 2006; Rossier, Meyer de Stadelhofen, & Berthoud, 2004). In these cases, it is likely that a successive integer scoring scheme was employed, in line with the recommendations of the test manual (Conn & Rieke, 1994). In addition, some studies using sum scores explicitly report using a successive integer scoring scheme (e.g., Booth & Irwing, 2011; Irwing, Booth, & Batey, 2014; these studies used summed item parcels). The use of one other scoring strategy was identified. Chernyshenko, Stark, and Chan (2001; see also Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001; discussed later) explored the hierarchical structure of the 16PF5 using dichotomized item responses that collapsed the *?* option into either the *a* or *c* options. The justification for this was based on the authors’ experience with analyzing the inventory and the lack of substantial influence of how the *?* is treated on previous findings. It is likely that the decision was also based on low endorsement rates for the middle response option and the consideration that modeling such data without collapsing this with another category would have led to unstable or uninterpretable parameter estimates.

However, the effects of dichotomization require further exploration and although it might have only a small distorting effect on sample statistics such as correlations, its effects could be more severe when estimating latent trait values for each individual in the sample. This might bias individual inferences in, for instance, clinical practice. Treating the *?* as missing—another strategy that we expected to find—was not to our knowledge used in any example of an empirical study using the 16PF5. Furthermore, there are two options when dichotomizing: to combine the middle response option with the lower response option or to combine it with the higher response option. Depending on the functioning of the middle response option, one of these strategies is likely to be more appropriate than the other and this could vary by subscale or even item. Fortunately, if researchers choose to dichotomize responses, the most appropriate way to do so for a given item is an empirical question answerable by examining how the middle option tends to be used by participants; however, it is a question that is as yet unanswered for the items of the 16PF5.

In fact, studies examining the 16PF5 at the level of the items have provided little evidence bearing on the functioning of the *?* option because all have made an a priori assumption about category ordering (Aluja & Blanch, 2004; Chernyshenko, Stark, Chan, et al. 2001; Ellis & Mead, 2000; Irwing et al., 2014). To our knowledge, only one study has provided direct evidence about the functioning of the middle response option in the 16PF5 (González-Romá & Espejo, 2003). They assessed the hypothesis that *?* lies between 0 and 1 in the 16PF5, using Bock’s (1972) nominal response model (NRM), but did so for a very limited set of items.

González-Romá and Espejo (2003) selected four items of the Social Boldness scale of the 16PF5 and studied the functioning of the middle response option. They found that for three items, the middle response option was not performing as an optimal indicator of intermediate levels of the latent trait because there was no interval of latent trait values for which it was the most probable response option. However, in terms of its location on the latent trait continuum, the *?* option did sit between the low and high response options for all items and so the more important assumption of correct category orderings was met. Although similar in intention to this study, the authors analyzed this small subset of items in isolation—not as part of the full 10-item Social Boldness scale from which they were selected; therefore, the implications for use of this scale and indeed the remainder of the 16PF5 in practice are not clear. Furthermore, they administered the items in the context of an experimental manipulation in which participants were offered the same items but with multiple different response formats. Specifically, participants saw the same set of items with four different response formats (*?* vs. *Not sure* vs. *In between* vs. no middle option), separated by a distracter set of questions. The order of presentation of the item sets was randomized, and as a result, many participants will have answered the same items with an alternative response format before answering the item with the current 16PF5 response format with ? as the middle option. It is very likely that this design and its obvious emphasis on the middle option affected responses to the 16PF5-format items. Therefore, the study is unlikely to have captured responding to the 16PF5 items as it occurs when the instrument is administered in practical applications or in other research settings.

Finally, the sample included a moderate number (n = 816) of undergraduate students. Ideally, studies of the performance of inventories should be conducted on large population-representative samples to improve the generalizability of findings.

Given the very limited amount of research conducted to date regarding the relative location of the 16PF5 middle response options on their respective latent trait continuums, it was our aim in this study to test whether the ostensibly middle response option really does have its assumed function as an indicator of intermediate levels of the latent trait. Extending the work of Gonzáles-Romá and Espejo (2003), we examined the functioning of the middle response option in the entire set of 158 items of the 16PF5 (excluding the Reasoning and Impression Management scales) in two large and independent standardization samples.

Method

Participants

16PF5 U.S. and UK standardization samples

We used the 16PF5 standardization samples from the United States (N = 10,261) and United Kingdom (N = 1,212) from a total sample of N = 10,261).[1] The U.S. standardization sample was reviewed in 2002 based on the U.S. Census in 2000 to ensure it remained representative of the general population of the United States with respect to a number of demographic variables including gender, ethnicity, age, and geographic region. The test publishers note that the educational level and years in education of the sample is greater than that of the U.S. population.

The UK version of the 16PF5 was initially standardized in 1993 through a joint data collection from ASE consultants and the Office of Population Censuses and Surveys (OPCS). The current sample was collected in 2011 by Oxford Psychologists Press (OPP) via an online market research panel. Participants received a small financial incentive for completing the questionnaire. Data were collected until quotas based on UK population demographics for sex, age, education, and level of employment were satisfied. The sample is largely representative of the underlying population with respect to age, gender, and years of education. Table 1 contains information on the demographic characteristics of each sample.

Table 1. Demographic characteristics of the U.S. and UK standardization samples of the 16PF5.

UK sample (N = 1,216)			U.S. sample (N = 10,261)		
	n	%		n	%
Gender			Gender		
Male	606	50.0%	Male	5,124	49.9%
Female	606	50.0%	Female	5,137	50.1%
Age (years)			Age (years)		
16–19	51	4.2%	15–24	3,714	36.2%
20–24	99	8.2%	25–44	4,282	41.7%
25–34	331	27.3%	45–54	1,614	15.7%
35–49	449	37.0%	55–64	577	5.6%
50–65	282	23.3%	65+	74	0.7%
Education level			Education level		
School-pre GCSE	47	5.1%	HS graduate or less	2,541	24.7%
School GCSE	219	23.6%	Some college	2,901	28.3%
A-level	194	20.9%	College graduate	4,819	47.0%
University (1st year)	55	5.9%			
University (2nd year)	100	10.8%			
Bachelor's degree	225	24.3%			
Master's degree	73	7.9%			
Doctorate	9	1.0%			
Postdoctorate	5	0.5%			
Employment status					
Full-time (self or other)	747	61.4%			
Part-time (self or other)	238	19.6%			
Unemployed	68	5.6%			
Student	77	6.3%			
Homemaker	43	3.5%			
Retired	34	2.8%			
Volunteer	5	0.4%			

Note. GCSE = General Certificate of Secondary Education.

Measures

16PF5

The 16PF5 contains 185 items organized into 15 primary personality scales each with between 10 and 14 items. The 15 primary personality scales are organized under five global (higher order) traits, namely Extraversion, which includes Self-Reliance (Q2), Warmth (A), Liveliness (F), Privateness (N), and Social Boldness (H); Anxiety, which includes Tension (Q4), Apprehension (O), Emotional Stability (C), and Vigilance (L); Tough-Mindedness, which includes Sensitivity (I), Openness to Change (Q1), Warmth (A), and Abstractness (M); Independence which includes Dominance (E), Social Boldness (H), Vigilance (L), and Openness to Change (Q1); and Self-Control, which includes Abstractness (M), Rule Consciousness (G), Perfectionism (Q3), and Liveliness (F).

In addition to the 15 primary personality scales, the 16PF5 includes a 15-item Reasoning scale and a 12-item Impression Management Scale that we did not use in this study. The response format for each of the items consists of a choice from three: a negative option for the trait, *?* and a positive option for the trait, scored as 0, 1, and 2 respectively. Respondents are instructed to avoid using the *?* option where possible.

The test manual for the 16PF5 reports internal consistencies for the 15 primary personality scales ranging from .66 to .86 (n = 4,460), test–retest reliability ranging from .69 to .87 over 2 weeks (n = 204), and .56 to .79 over 2 months (n = 159; Conn & Rieke, 1994, p. 81). Further, in a recent study, Irwing et al. (2014) found reasonable to good fit for single-factor,

[1] Copyright © 1993 by the Institute for Personality and Ability Testing, Inc., Champaign, Illinois, USA. All rights reserved. Reproduced with special permission of the publisher. Further reproduction is prohibited without permission of IPAT Inc., a wholly owned subsidiary of OPP Ltd., Oxford, England.

item-level confirmatory factor analysis (CFA) models of each of the primary scales based on the same data as this study (root mean square error of approximation range = .05−.09; comparative fit index range = .95−.99).

Statistical Procedure

Nominal response models

Originally developed as a means to study the functioning of distracters in multiple choice tests, the NRM can be used to empirically assess category ordering hypotheses because it places no constraints on the ordering of response options. The NRM is a very general model from the "divide by total" family of item-response theory models. Other familiar and commonly used divide by total models such as the generalized partial credit model (GPCM; Muraki, 1992) and partial credit model (Masters, 1982) can be obtained from the NRM with the addition of appropriate model constraints (see Preston et al., 2011). Note, however, that these and other models that place constraints on the ordering of categories cannot be used to provide a direct empirical test of category ordering because a specific category ordering is assumed a priori. A comprehensive discussion of the NRM, its properties, and its uses can be found in Mellenbergh (1995) and Preston and Reise (2014), which we draw on in our later description of the model.

In the NRM, the probability of an individual i with trait level θ_i endorsing response option $x = 0, \ldots m_k$ can be written as:

$$P_{kx}\ (\theta_i) = \frac{\exp(c_{kx} - a_{kx}\theta_i)}{\sum_{x=1}^{m}\exp(c_{kx} - a_{kx}\theta_i)} \tag{1}$$

The a_{kx} parameter is a slope parameter or category discrimination parameter that represents the rate of change in log-odds of responding in a response category with θ_i (Preston & Reise, 2014). The c_{kx} parameters are intercept parameters that reflect the relative popularity of a particular response category. Larger c_{kx} parameters indicate a more popular response option. For identification purposes, some parameters must be constrained. Bock (1972) suggested constraining the sum of $\mathrm{a_{kx}}$ and $\mathrm{c_{kx}}$, the parameters within an item, to zero. In most contemporary uses of the NRM, however, the $\mathrm{a_{kx}}$ and $\mathrm{c_{kx}}$ parameters of the first or last category are set to zero to achieve identification.

In the context of testing category ordering, it is interesting to consider the probability of choosing between one of two response categories assumed to be adjacent, x and $x' = x - 1$ (Thissen, Steinberg, & Fitzpatrick, 1989):

$$P_{kx} \mid x \ = \ x \ or \ \ x' = \frac{1}{1 + \exp\left(-a_k^{*}\theta + d_k\right)}$$

where a_k^* is $a_x - a'_x$ and $d_k = \ c'_x - c_x$. When a_k^* is positive, the probability of responding in category x versus x' increases as the level of θ increases. As a_k^* is positive when $a_x > \ a'_x$, the category response functions can be ordered based on their a_{kx} parameters. Therefore, the ordering of categories can be assessed by examining whether the a_{kx} parameters increase for successive response categories and whether all the corresponding a_k^* parameters are positive (Samejima, 1972, 1996; González-Romá & Espejo, 2003).

Second, it is possible to assess whether the middle option is a genuine middle option in the sense that there is an interval of the latent trait continuum where individuals with latent trait values in that range are more likely to select the middle response option than the response options on either side. This can be assessed by examining the response category thresholds; that is, the point on the latent trait continuum at which the category response curves for two response options intersect and at which, therefore, there is an equal probability of responding in either of the two categories. The thresholds between two response categories can be estimated from the parameters of the NRM as (González-Romá & Espejo, 2003):

$$\tau_{x,x-1} = \frac{c_{kx} - c_{k,\ x-1}}{a_{k,x-1} - a_{kx}} \tag{2}$$

where c_{kx} and $c_{\ k,x-1}$ are the location parameters for two response categories and a_{kx} and $a_{k,x-1}$ are the corresponding discrimination parameters. For the *?* option to be performing as a middle category, the threshold dividing the lowest and middle category τ_{01} should occur at a lower value of the latent trait than the threshold dividing the middle and highest category τ_{12}. The NRM, therefore, allows the testing of two key hypotheses relating to the functioning of middle options: that of ordered response options and of ordered thresholds.

Visual inspection of the category response curves for an item can also help diagnose unexpected functioning of the middle category option. When the middle option is functioning as implied by the numerical values assigned to the responses, then plotting the category response curves for the three response options would result in a plot similar to that in Figure 1. In Figure 1, the category response curves are in the expected order based on their assigned numerical values in terms of their location along the latent trait continuum. That is, the category response curve for the 0 option is located to the left of that of the 1 option, which is in turn located to left of the 2 option. In addition, each response option is the option most likely to be selected by respondents for a given interval of the latent trait; that is, there is a range of latent trait values for which the height of the response curves for 0, 1, and 2 exceed that of the other two response curves. For the 0 option this is at lower levels of the latent trait, for the 1 option this is at intermediate levels of

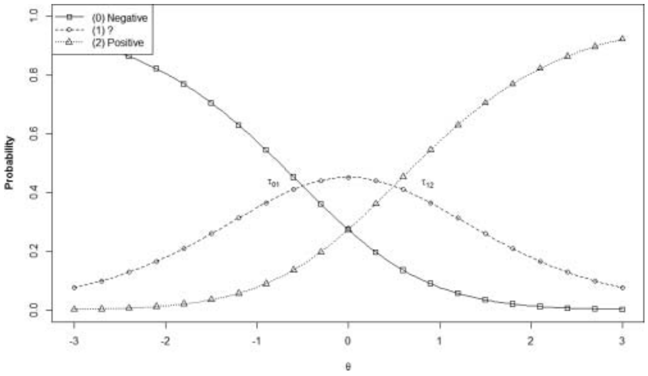


Figure 1. Example category response curves with thresholds and discriminations in the correct order.

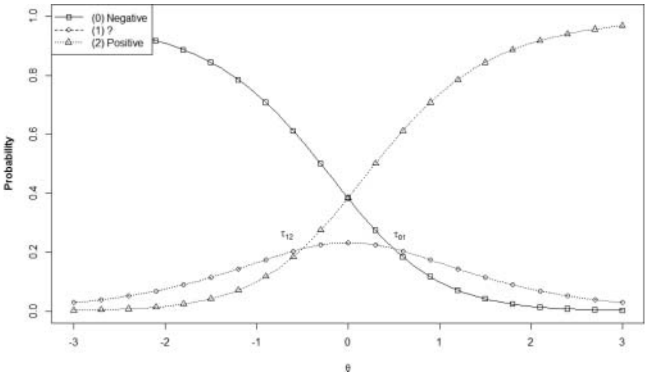


Figure 2. Example category response curves with thresholds disordering.

the latent trait, and for the 2 option this is at higher levels of the latent trait.

Deviations from this pattern would imply that the middle option is not functioning as is typically assumed (i.e., when using a successive integer scoring scheme). There are two ways in which this can occur: category disordering and category threshold disordering. Figure 2 illustrates category threshold disordering. Here, the category response curves are in the order that corresponds with their assigned numerical values but there is no interval of the latent trait for which the height of its category response curve exceeds that of the category response curves for the lowest and highest response option. This means that the 1 option is never the most likely response option, even for middle values of the latent trait. Category disordering is illustrated in Figure 3. Here, the category response curve for the middle option lies to the left of that for the response option with the lowest assigned numerical value. This means that the categories are in the correct order and selecting 1 is actually indicative of lower levels of the latent trait than selecting 0.

Models were estimated using the mcIRT package (Reif, 2014) in the R statistical software package (R Core Team, 2013). The package uses the parameterization given in Equation 1. For identification purposes, the a_{kx} and c_{kx} parameters for the last response option were constrained to zero for each item.

Item evaluation

a_{kx}* *Parameters. In this study, where the a_{i2} is fixed to zero for each item, for response categories to show the correct ordering, a_{k0} and a_{k1} should both be negative with a_{k0} being more negative than a_{k1}. We examined a_{kx} parameters and noted for which items these occurred in an unexpected order. We distinguished between cases in which the middle option behaved as though it was the lowest response option and between cases in which the middle option behaved as though it was the highest response option.

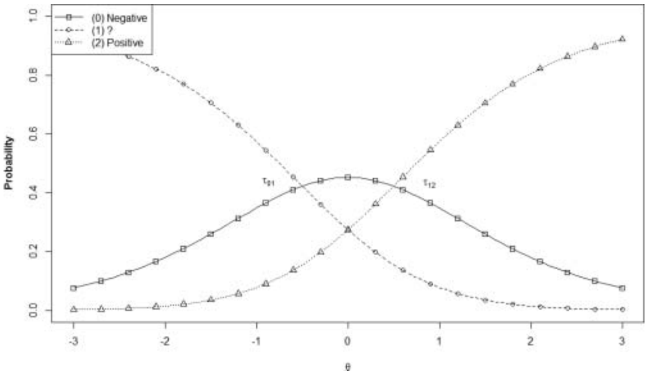


Figure 3. Example category response curves with discriminations disordering.

Category thresholds. The thresholds between two categories were computed from the parameter estimates of the NRM as in Equation 2 earlier. These thresholds represent the point on the latent trait continuum at which an individual has equal probability of responding in either of the two adjacent categories. For the *?* option to be performing as a genuine middle category, the threshold dividing the lowest and middle category should occur at a lower value of the latent trait than the threshold dividing the middle and highest category. Category threshold disordering in the context of ordered a_{kx} parameters implies that there is no interval of the latent trait for which the middle response option is the most likely to be selected. In cases where response categories are in the incorrect order, we do not interpret the ordering of the thresholds because the category ordering is a more fundamental assumption than category threshold ordering. Although the disordering of category thresholds is not as problematic as disordering of the a_{kx} parameters, it still suggests that the middle option is not functioning as intended because a genuine middle option should be the most likely response option for middle values of the latent trait.

Statistical tests

In addition to the descriptive analyses of response category ordering described earlier, we also compared the fit of the NRM to that of the GPCM (Muraki, 1992). The GPCM is nested within the NRM. In the GPCM, the ordering of categories is constrained because the a_{kx} parameters are fixed to x, where x is the integer assigned to a response option according to a consecutive integer scoring scheme. This guarantees the ordering of the response options. A comparison of the fit of the GPCM to the NRM, therefore, provides a further test of the category ordering hypothesis. We compared the fits of the GPCM and NRM based on $\Delta\chi^2$, ΔAIC, ΔBIC, and $\Delta saBIC$. We considered the difference in fit to be practically significant when ΔBIC was >10 (Raftery, 1995). Models were estimated using the 'mirt' package in R statistical software (Chalmers, 2012; R Core Team, 2013).

Results

Response scale evaluation: U.S. standardization sample

Table 2 provides information on number of items, endorsement rates for the middle response category, and summary information about the performance of the middle option for each primary scale. Provided is information on the following:

1. How many items had response options that were not correctly ordered.
2. How many items with correctly ordered response options had category thresholds that were not correctly ordered.

Table 2. Nominal response model (NRM) model results for U.S. standardization sample.

				Results from NRM				NRM versus GPCM fit comparisons			
Scale	Label	No. items	Mean ? endorsement rate	Items where ? behaves as lowest response option	Items where ? behaves as highest response option	Items with thresholds out of order	$\Delta\chi^2$	*df*	ΔAIC	ΔBIC	$\Delta saBIC$
A	Warmth	11	.08	8	—	3	560.27	11	−538.27	−458.67	−493.63
C	Emotional Stability	10	.07	—	—	10	270.81	10	−250.81	−178.45	−210.23
E	Dominance	10	.09	—	—	10	138.90	10	−118.90	−46.54	−78.31
F	Liveliness	10	.10	—	—	10	59.78	10	−39.78	32.58	0.80
G	Rule Consciousness	11	.09	—	—	11	455.23	11	−433.23	−353.63	−388.59
H	Social Boldness	10	.09	—	—	10	55.15	10	−35.15	37.21	5.43
I	Sensitivity	11	.10	—	—	11	40.75	11	−18.75	60.85	25.89
L	Vigilance	10	.14	—	—	10	30.72	10	−10.72	61.65	29.87
M	Abstractness	11	.12	—	—	11	146.77	11	−124.77	−45.17	−80.13
N	Privateness	10	.09	—	—	10	38.16	10	−18.16	54.20	22.42
O	Apprehension	10	.08	—	—	10	118.61	10	−98.61	−26.25	−58.03
Q1	Openness to Change	14	.10	14	—	—	619.47	14	−591.47	−490.17	−534.66
Q2	Self-Reliance	10	.11	—	—	10	77.36	10	−57.34	15.01	−16.77
Q3	Perfectionism	10	.07	—	—	10	66.55	10	−46.55	25.82	−5.96
Q4	Tension	10	.09	—	—	10	40.73	10	−20.73	51.64	19.86

Note. GPCM = generalized partial credit model; AIC = Akaike's information criterion; BIC = Bayesian information criterion. Negative values of ΔAIC, ΔBIC, and $\Delta saBIC$ indicate that the NRM fits better than the GPCM.

When the category response options were in the incorrect order, we note when the middle option behaved as the lowest response option versus the highest response option.

The average middle response option endorsement rates for a scale ranged from 7% in the C (Emotional Stability) and Q3 (Perfectionism) scales up to 14% in the L (Vigilance) scale. With the exception of the Q1 (Openness to Change) scale, most items showed discrimination parameters that were in the correct order. For Q1, the *?* option was functioning as the lowest response option rather than as the middle response option. Other items that showed incorrect ordering of categories were Items 1, 4, 5, 7, 8, 9, 10, and 11 from the A scale (Warmth). In these items, as in the Q1 scale, low rather than intermediate levels of latent trait were associated with selecting the *?* response option. Neither of these scales showed anomalous rates of middle category endorsement as compared to the other scales of the 16PF5. Among the items in which the response categories were technically in the correct order, there were many cases in which the discrimination parameter of the middle response category was very similar to that of an adjacent category. That is, the strength of ordering of the response categories was poor. For example, in Item 6 of the A (Warmth) scale, a_{i0} and a_{i1} were −0.99 and −0.95, respectively, suggesting that these categories were not clearly distinguishable in terms of their location on the latent trait continuum.

Further, among those items that showed correctly ordered discrimination parameters, none showed category thresholds that were in the correct order. Therefore, there were no items in which the *?* option was performing as a genuine middle option. Overall, the majority of items had category response curves that resembled those in Figure 2 in which discrimination parameters but not the category thresholds were in the correct order. This is consistent with the relatively low endorsement rates of the *?* option, which can result in this response option never being the most probable response option. Under these circumstances, the thresholds will be in the incorrect order.

The comparisons of model fit for the NRM against the GPCM in the U.S. standardization sample also suggested that category ordering was not always as assumed. In the A (Warmth), O (Apprehension), C (Emotional Stability), L (Sensitivity), Q1 (Openness to Change), M (Abstractness), G (Rule Consciousness), and E (Dominance) scales, the NRM was better fitting to a level that could be considered practically significant according to ΔBIC > 10. The difference in fit favoring the NRM was most pronounced for those scales in which the NRM parameter estimates identified out-of-order response categories (Q1 and A) but was still substantial for many others. In the G scale, for example, in which fit indexes also strongly favored the NRM, the discrimination parameters were ordered but not strongly so. Specifically, the a_{k0} parameters were close in magnitude to corresponding a_{k1} parameters within each item (e.g., −1.27 vs. −1.15 for Item 5). In these cases, the middle option, although technically falling in the middle of the response scale-, was difficult to distinguish from the lowest response option.

Response scale evaluation: UK standardization sample

Results analogous to those in the U.S. standardization sample for the UK sample are provided in Table 3. The average middle category response option rates for this scale were higher than those for the U.S. standardization sample and ranged from 11% for the Q4 (Tension) scale up to 19% for the M (Abstractness) scale. These are also higher than middle category endorsement rates of between 5% and 10% reported in previous research (Chernyshenko, Stark, Chan, et al., 2001). There were a large number of items that did not show the correct ordering of discrimination parameters. Like the U.S. sample, this included the entire Q1 scale (Openness to Change) in which the *?* response option functioned as the lowest response option. Other items in which the discrimination parameters were in the incorrect order were several items from Q4 (Tension), one item from N (Privateness), one item from A (Warmth) and all items from L (Vigilance), G (Rule Consciousness), and O (Apprehension). In some cases, the *?* option was functioning as the lowest response option (all items of G [Rule Consciousness]), whereas in others it was functioning as the highest response option (Items 4, 7, and 8 of Q4 [Tension];

Table 3. Item response theory model results for UK standardization sample.

				Results from NRM					NRM versus GPCM model fit comparisons	
Scale	Label	No. items	Mean *?* endorsement rate	Items where *?* behaves as lowest response option	Items where *?* behaves as highest response option	Items with thresholds out of order	$\Delta\chi^2$	*df*	ΔAIC	ΔBIC	$\Delta saBIC$
A	Warmth	11	.13	—	1	10	41.71	11	−19.71	36.40	1.456
C	Emotional Stability	10	.14	—	—	10	15.71	10	4.29	55.29	23.53
E	Dominance	10	.14	—	—	10	20.54	10	−0.54	50.46	18.70
F	Liveliness	10	.15	—	—	10	14.52	10	5.48	56.48	24.72
G	Rule Consciousness	11	.12	11	—	—	455.23	11	−433.23	−353.63	−388.59
H	Social Boldness	10	.13	—	—	10	19.13	10	0.87	51.87	20.10
I	Sensitivity	11	.13	—	—	11	40.75	11	−18.75	60.85	25.89
L	Vigilance	10	.18	—	10	—	261.59	10	−241.56	−190.56	−222.34
M	Abstractness	11	.19	—	—	11	146.77	11	−124.77	−45.17	−80.13
N	Privateness	10	.13	—	1	9	59.36	10	−39.36	11.64	−20.12
O	Apprehension	10	.13	—	10	—	276.24	10	−256.24	−205.24	−237.01
Q1	Openness to Change	14	.14	14	—	—	645.22	14	−617.21	−545.81	−590.28
Q2	Self-Reliance	10	.13	—	—	10	17.53	10	2.47	53.47	21.71
Q3	Perfectionism	10	.12	—	—	10	38.79	10	−18.79	32.22	0.45
Q4	Tension	10	.11	—	3	7	60.09	10	−40.09	10.91	−20.86

Note. NRM = nominal response model; GPCM = generalized partial credit model; AIC = Akaike's information criterion; BIC = Bayesian information criterion. Negative values of ΔAIC, ΔBIC, and $\Delta saBIC$ indicate that the NRM fits better than the GPCM.

Item 7 of N [Privateness]; Item 1 of A [Warmth]; all items of L [Vigilance]; all items of O [Apprehension]). There were no items that showed both discrimination parameters and category thresholds that were in the correct order.

The comparisons of model fit for the NRM against the GPCM in the UK standardization sample provided further evidence for unexpected response behavior regarding the *?* option. Based on our criteria of ΔBIC > 10, the fit of the NRM was favored to a practically significant level in four scales: O (Apprehension), L (Vigilance), Q1 (Openness to Change), and G (Rule Consciousness). All of these scales were identified as having a large number of items with category response options that were out of order. Our model fit criterion did not identify the cases previously identified by inspection of NRM parameter estimates when only a small number of items in a scale had response categories in the correct order.

Illustrative examples

To appreciate what these results mean, we discuss two illustrative examples in depth. First, consider Item 1 from the Sensitivity (I) scale in the U.S. standardization sample, which showed category responses in the correct order but thresholds in the incorrect order. The category endorsement rate for the middle option for this item was 10%. The NRM parameter estimates for this item are provided in Table 4. It can be seen that the a_{kx} parameters are correctly ordered (the discrimination of the first category is more negative than the second category); therefore, the response categories are also in the correct order. However, the category thresholds are in the incorrect order. That is, the category response curve for the middle response option intersects with the category response curve for the highest response option at a lower latent trait level than it intersects with the category response curve for the lowest response option. This feature can be seen in Figure 4, which plots the category response curves for the item. Note that at no point along the latent continuum is the middle response option the most probable response option.

Another illustrative example is Item 1 from the Openness to Experience (Q1) scale from the U.S. standardization sample. In this instance, the response categories were in the incorrect order. Parameter estimates for this item are also provided in Table 4. For this item, the a_{kx} values place the*?* not in the middle of the response scale, but further down the latent trait continuum than the ostensibly lowest response option. That is, a_{k1} is more negative than a_{k0} (recall that a_{k2} is fixed to zero for identification). This incorrect ordering of categories can be seen in the category response curves for the item in Figure 5.

Practical implications of category disordering

Finally, to investigate the implications of disordering of categories, we estimated scores for the Q1 scale in the 16PF5 U.S. standardization sample using the following:

1. NRM factor scores estimated from the NRM factor scores expected using a posteriori method (Embretson & Reise, 2000).
2. A simple summing of item scores ("sum scores").
3. Sum scores treating the middle option as missing and using the mean of the remaining scores as the trait estimate.

Table 4. Parameter estimates from nominal response model for two illustrative examples.

Item	a_{k0} (SE)	a_{k1} (SE)	a_{k2} (SE)	c_{k0} (SE)	c_{k1} (SE)	c_{k2} (SE)	τ_{01}	τ_{12}
I Item 1	−1.30 (0.03)	−0.67 (0.05)	0.00[a] (N/A)	0.35 (0.02)	−1.49 (0.04)	0.00[a] (N/A)	2.92	−2.22
Q1 Item 1	−0.57 (0.03)	−1.29 (0.05)	0.00[a] (N/A)	−1.04 (0.02)	−3.01 (0.06)	0.00[a] (N/A)	−2.74	−2.33

[a] Fixed for identification purposes.

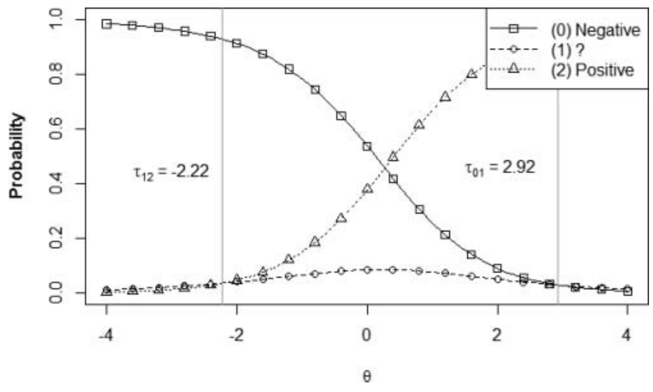


Figure 4. Category response curves for item 1 of the Sensitivity (I) scale in the U.S. standardization sample.

4. Sum scores recoding the middle option as 0; that is, collapsing the category with the lowest response option.
5. Sum scores recoding the middle option as 2; that is, collapsing the category with the highest response option.

We then used these scores in two different types of analyses to mimic research and real-world applications. First, we estimated a correlation with another trait; second, we considered rank ordering of respondents in making a selection decision. We chose the Q1 scale because in exhibiting category disordering for all 14 items, it could act as a worst case scenario.

Figure 6 plots each of the Q1 sum scores (y axes) and the NRM factor scores (x axis). All scores have been standardized to have a mean of 0 and a variance of 1 to facilitate comparison. A reference line shows where the sum scores would be expected to sit, were they identical to the NRM factor scores. The correlation between the factor scores and the simple sum scores was relatively high at $r = .83$, lower when using the missing data strategy at $r = .77$, lower still when collapsing the middle response category with the top category at $r = .62$, and highest when collapsing the middle response category with the lowest response category at $r = .95$. This is consistent with the fact that the NRM results suggested that selecting the middle category was informative about trait levels (making the missing data strategy suboptimal); however, it indicated low, rather than intermediate trait levels (making collapsing it with the lowest category superior to collapsing it with the highest response category).

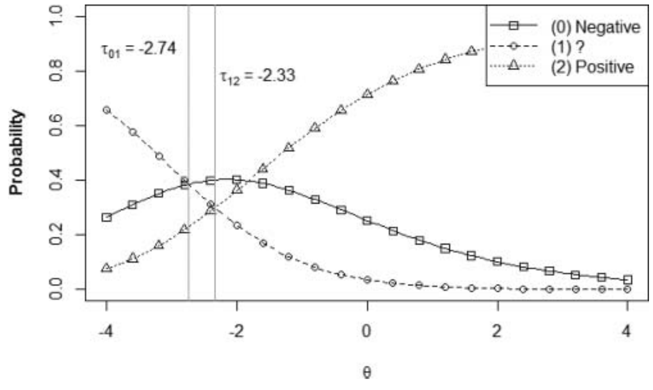


Figure 5. Category response curves for Item 1 of the Openness to Change (Q1) scale in the U.S. standardization sample.

To investigate the impact of scoring method on criterion correlations, we computed correlations between Q1 NRM factor scores and Q1 sum scores with the 16PF5 Reasoning scale scores. The Reasoning scale is a short (15-item) cognitive ability measure assumed to assess verbal, numerical, and logical abilities. It is designed to provide a quick measure of intelligence and is moderately correlated ($r = .61$) with the Information Inventory and with the Form A, Scale 2 Culture Fair Intelligence Test ($r = .51$; Conn & Rieke, 1994). The test manual reports a Cronbach's alpha of .80 for the scale with 2-week and 2-month test−retest reliabilities of .71 and .70, respectively. Investigations of differential item functioning by gender and ethnicity found no biasing by race or gender, with the exception of one item that functioned differently in a Hispanic sample (Conn & Rieke, 1994).

The correlation between Reasoning scores and NRM scores was $r = .20$, with the simple sum scores it was $r = .21$, with the sum scores using a simple missing data strategy it was $r = .21$, with the sum scores collapsing the middle response category with the highest response category it was $r = .19$, and with the sum scores collapsing the middle response category with the lowest response category it was $r = .21$. The method of scoring, therefore, made very little difference to a criterion association.

Finally, to mimic the use of the test in a selection setting we selected the top 100, 500, and 1,000 participants based on NRM factor scores and compared it to selection based on the three kinds of sum scores. We assessed the extent to which the same participants were selected depending on the scoring method. Results are provided in Table 5 and are the percentage of individuals selected according to the NRM factor scores who were also selected according the various sum score estimates. When selecting the top 100, 500, and 1,000 participants, all sum score estimates tended to select a group of participants with only low to moderate overlap with those selected based on NRM scores. There was more overlap when selection was based on the bottom 100, 500, and 1,000 participants. Overall, collapsing the middle category with the lowest response category showed the most similar selections to the NRM factor scores and collapsing with the highest response category showed the worst. Again, this is consistent with the fact that the use of middle response category was indicative of low trait levels according to the NRM results.

Summary

In all, 22 items out of the 158 items of the 16PF5 that were evaluated in this study had response categories that were in the incorrect order in the U.S. sample and 50 had response categories that were in the incorrect order in the UK sample. The Q1 scale was problematic in both samples, with the middle response option functioning as the lowest response option in both. In addition, none of the 158 items had both response categories and category thresholds in the correct order in either sample. The practical implications of category disordering were negligible for criterion correlations but significant for a selection decision. This is consistent with the idea that category disordering might matter little for group-level analysis but

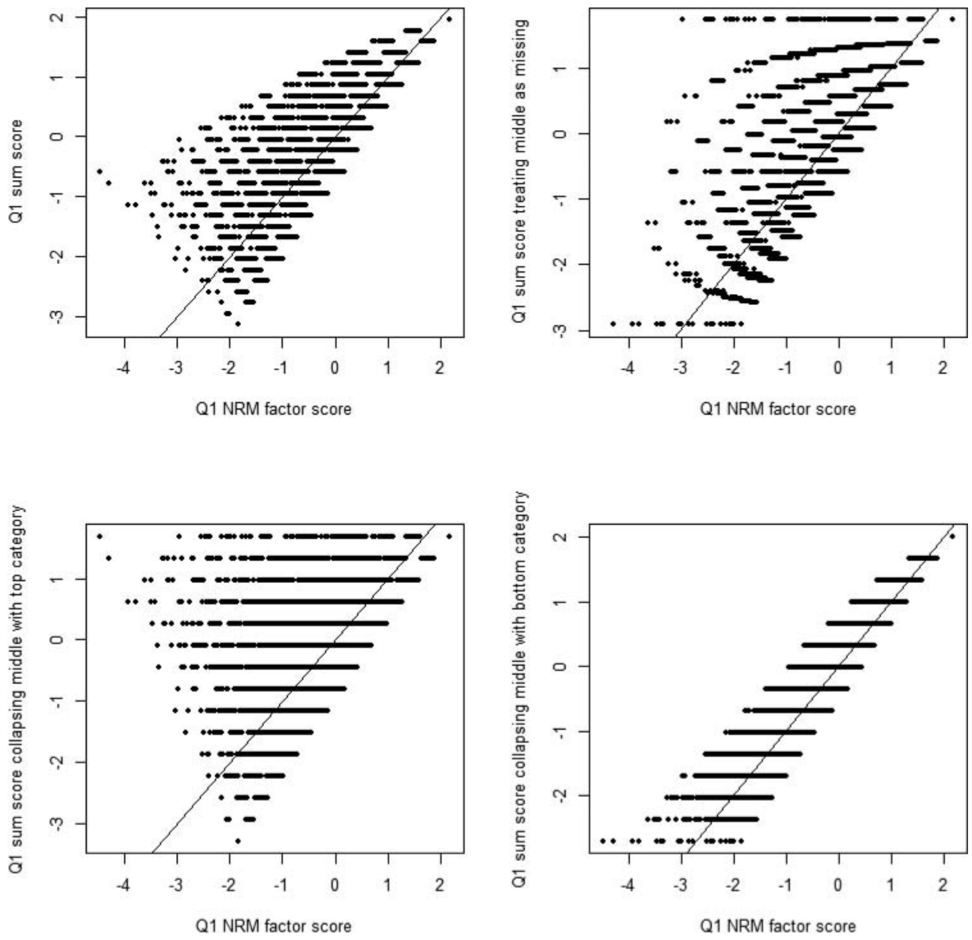


Figure 6. Nominal response model (NRM) factor scores plotted against different types of sum score for Openness to Change (Q1).

becomes important when scores are used for individual-level assessments and decisions.

Discussion

In this study, we found evidence that the middle option (*?*) of the 16PF5 personality scales does not always behave as an indicator of intermediate levels of the latent trait as implied by the consecutive integer scoring scheme usually adopted to score the items in this inventory. In many cases, the middle response option was actually associated with levels of the latent trait lower than the ostensibly lowest response option, and in others it was associated with levels of the latent trait higher than the ostensibly highest response option. In such instances, using consecutive integers to score the low, middle, and high response options is not appropriate and will lead to distorted estimates of the latent trait. A secondary observation was that even when response categories were in the order implied by their typically assigned numerical values, category thresholds were not correctly ordered. This meant that there was no interval of the latent trait for which the middle response option was the most probable response. Although this is less serious than response options being in the incorrect order, it reflects a lack of use of the middle response option and thus calls into question its utility. In terms of the practical implications of unintended use of the middle response option, we showed that simulating a selection scenario for the worst-affected scale, the use of sum scores versus factor scores estimated according to the NRM led to widely diverging selection decisions. The use of sum scores as trait estimates for high-stakes selection decisions is, therefore, potentially problematic when middle response category use is not as assumed. However, criterion correlations were essentially invariant to scoring method, suggesting that the scoring method will matter little in many research settings.

Table 5. Percentages of participants selected by nominal response model factor scores who were also selected by simpler scoring methods.

Selection criterion	Sum score	Treat middle category as missing	Collapse middle category with lowest category	Collapse middle category with highest category
Top 100	2%	16%	29%	0%
Top 500	25%	35%	55%	14%
Top 1,000	43%	52%	69%	30%
Bottom 100	100%	45%	100%	42%
Bottom 500	69%	43%	80%	38%
Bottom 1,000	71%	65%	82%	48%

Results highlight the need to empirically assess the functioning of response categories, rather than assuming that they function in the manner implied by the numerical scores assigned to response options. This can be done following the procedure in this study and other similar studies (Gonzáles-

Romá & Espejo, 2003). Although the middle response category of the 16PF5 might represent a particularly ambiguous response option, other personality inventories use middle response options with verbal labels that are subject to at least some degree of ambiguity of interpretation. For example, the middle option of three commonly used personality inventories are *?*(16PF5; Conn & Rieke, 1994), *neither accurate nor inaccurate* (the items included in the International Personality Item Pool; Goldberg, 1999), and *neutral* (NEO Five-Factor Inventory; Costa & McCrae, 1992). Such response options are also liable to elicit responses that are due to factors other than being of an intermediate level of the latent trait. Therefore, the need to assess the functioning of response categories also applies to other commonly used personality inventories.

In particular, we recommend a combined strategy of examining NRM parameters and comparing model fits with the GPCM. The former can identify individual problematic items and the latter provides a more objective statistical test regarding the response options across items in the scale as a whole. Substantive researchers have several software options including the freely available 'mcIRT' (Reif, 2014) or 'mirt' packages in the R statistical software package (Chalmers, 2012), or 'MIRT' software (Glas, 2010).

To date, few previous studies have used the NRM and nested models to study category responding; however, in those studies that have, evidence generally suggests that response categories but not response category thresholds tend to be in the correct order. For example, Hernández, Espejo, González-Romá, and Gómez-Benito (2001) studied the performance of a 5-point Likert scale in three measures of job satisfaction, role overload, and support climate, respectively, with an ambiguous middle option labeled *indifferent.* They found that although the response options were in the correct order, for several items the category thresholds were not. González-Romá and Espejo (2003), as described in detail earlier, found similar results in a small subset of 16PF5 items. However, as this study illustrates, category disordering will also sometimes arise and category ordering should be tested, not simply assumed.

In cases such as this study in which the categories appear in the incorrect order, the most defensible solution is to obtain latent trait estimates using a measurement model such as the NRM used here that takes category disordering into account. More crude strategies such as scoring the middle option as missing or combining it with another response option are suboptimal because, like simple consecutive integer scoring, they also entail an implicit scoring scheme with assumptions that are likely to be violated under disordered response categories or thresholds. For example, when combining the middle option with another response category, this makes an implicit assumption that the *?* response carries the same information about the latent trait level as responding in the category with which it is combined. This is in contrast to the scoring scheme engendered in the NRM, in which the contribution of category response to latent trait estimates are determined by category discrimination parameters and thus take into account the differential information conferred by responses in different categories.

We evaluated the potential implications of using sum scores in practice when the middle response category is not used in the manner assumed using Q1, the worst affected scale in this respect. We found that criterion correlations are little affected by whether a sum score or a factor score from the NRM is used, suggesting that parameter estimates based on aggregated data are unlikely to be strongly affected. This would suggest that the use of affected scales in research settings is largely unproblematic.

However, when we simulated selection based on various kinds of sum scores and compared this to selection based on factor scores estimated from the NRM, we found that there was little to moderate overlap in the individuals selected by the two alternative methods. Although there will be other sources of unreliability that will have contributed to the discrepancies between the two scoring methods, an important characteristic of category disordering is that it led on average to an overestimation of trait levels for individuals at the lowest levels when using sum scores relative to NRM scores. These individuals had a tendency to select the middle response option, earning them an item score of 1 when in fact a more appropriate item score would have been 0.

Of the kinds of sum scores compared, the best strategy was to collapse the middle category with the response category to which the NRM results suggested it was closest. In the Q1 scale, the middle response option was behaving as if it were the lowest response category. As a result, the best results were obtained when the middle and lowest response categories were collapsed and the worst results were obtained when the middle and highest response categories were collapsed. Treating the middle category as being in between the lowest and highest response categories and treating endorsements of the middle response category as missing data yielded results intermediate between these two extremes.

These results suggest that the best strategy after estimating factor scores would be to first establish which response category the middle response category is closest to and then collapse it with this category, but that this depends on the item in question. For the 16PF5, the current results would suggest that for most scales, the middle response category should continue to be scored as a middle response. In U.S. samples, exceptions are all items of the Q1 scale and Items 1, 4, 5, 7, 8, 9, 10, and 11 of the A scale where the middle response category should be collapsed with the lowest response category. In UK samples, exceptions are all items of the Q1 scale and all items of the G scale where the middle response should be collapsed with the lowest response option and Items 4, 7, and 8 of Q4; Item 7 of N; Item 1 of A, and all items of L and O, where the middle response option should be collapsed with the highest response option.

We also noted that, although it was only the minority of items that showed out-of-order response options, all of the remaining items showed out-of-order category thresholds. This means that there is no value of the latent trait for which the middle response option is the most probable response. We do not, however, think that this observation is problematic because it does not imply any systematic bias in scoring. It most likely reflects the low endorsement rates of the middle response option and is not surprising given that the test instructions encourage respondents to avoid using the middle response option where possible. At worst, it suggests that the middle response option is uninformative about latent trait levels and is thus superfluous. However, amending test instructions such that respondents felt free to use the middle response options would most likely reverse this situation and result in category thresholds that were in the correct order.

In terms of the implications for future test construction, our results highlight the challenge of providing participants with a middle response option—namely, that it will tend to elicit responses other than indications that an individual possesses intermediate trait levels. In spite of this, it has been suggested that middle response options could be important for minimizing missing data and maintaining participant cooperation (e.g., Rammstedt & Krebs, 2007). Therefore, we would argue that middle response options can be useful; however, any ambiguity in the associated verbal labels should be carefully avoided. If the middle response option is intended to function as a marker of intermediate trait levels, a verbal label such as *In between* should be used, rather than *Unsure* or *?*. Similarly, any other factors that have been suggested to promote unintended use of the middle category such as poor item comprehensibility (promoting confusion), or a lack of contextualization of items (promoting "it depends" responses) could be minimized to the extent that this is appropriate in the context of a given measure (Kulas & Stachowski, 2009). In addition, participants could be offered an additional unscored response category such as a *Not applicable* option to help avoid unintended uses of the middle category (Kulas et al., 2008).

As well as highlighting the need to assess category ordering empirically, our results are also potentially informative about the role of personality itself in influencing the manner in which individuals tend to use response scales. Evidence from both samples suggested that there was an increased likelihood of selecting the *?* option in individuals who were low in openness to change. One possibility is that individuals who are low in openness to change are more intolerant of ambiguity or uncertainty and thus select the *?* option when not entirely certain about their response, rather than selecting a closest option from the 0 or 2 options. In addition, there was some evidence from the UK sample that individuals were more likely to select ? when low in rule consciousness. This same general tendency was observed in the U.S. sample, in which the low and *?* response options were difficult to distinguish (highly similar discrimination parameters) but the tendency was not so marked as to result in a reversal of the two response categories. This could reflect a greater willingness to ignore the test instruction that requests that respondents aim to avoid selecting the *?* response.

Neurotic traits were also associated with selecting the *?* response: The traits of apprehension, vigilance, and tension all showed a positive relation with selecting *?* in the UK sample. This is consistent with previous studies reporting that individuals higher in neurotic traits are more likely to select a middle option (Hernández et al. 2004; Kulas & Stachowski, 2013; McFadden & Krug, 1984). Hernández et al. (2004), whose results are particularly relevant to those of this study in having used the 16PF5, proposed that this result could be explained by individuals who are higher in neuroticism having greater difficulty in choosing between the high and low response options and thus electing to avoid making the decision altogether by selecting the middle option. Consistent with our analyses of the U.S. sample, Hernández et al. (2004) also found that low warmth was associated with increased use of the *?* category and suggested that those low in warmth might be more reluctant to share personal information with others. Finally, we found that one item from the Privateness scale was associated with the *?* option in the UK sample, suggesting that participants with higher levels of privateness are less likely to disclose information about themselves, preferring to select *?*.

The preceding information highlights that although the Q1 scale showed the same pattern of *?* responding across U.S. and UK samples, there were also some differences between the results of the two samples. Overall, there were more items with out-of-order categories in the UK sample, which could reflect the reduced cultural relevance of such items for a UK respondent given that the scale was originally developed in the United States. As a result, UK respondents might have more difficulty interpreting items or feel that the appropriate response is *not applicable*, exacerbating any tendencies to select the *?* option. Indeed, the use of the middle response category was overall greater in the UK sample than in the U.S. sample. However, the differences between samples could simply reflect the fact that the size of the UK sample was smaller. Given that for many items, discrimination parameters for the middle and an adjacent category were very close in value, this would lead to these categories being out of order by chance more often in the UK sample.

Recommendations for scoring the 16PF5 in practice

Based on our results it is possible to recommend the following scoring scheme for the use of the 16PF5 in practice. If administering the 16PF5 to a U.S. sample, the following scoring scheme should be used:

- Use $a = 0$, $? = 1$, and $c = 2$ for all items in scales C, E, F, G, H, I, L, M, N, Q2, Q3, and Q4, and Items 2, 3, and 6 of scale A.
- Use $a = 0$, $? = 0$, and $c = 2$ for all items in scale Q1 and for Items 1, 4, 5, 7, 8, 9, 10, and 11 of scale A.
- Do not use $a = 0$, $? = 2$, and $c = 2$ as a scoring scheme for any item.

If administering the 16PF5 to a UK sample, the following scoring scheme should be used:

- Use $a = 0$, $? = 1$, and $c = 2$ for all items in scales C, E, F, H, I, M, Q2, and Q3; Items 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, and 11 of the A scale; Items 1, 2, 3, 4, 5, 6, 8, 9, and 10 of the N scale; and Items 1, 2, 3, 5, 6, 9, and 10 of the Q4 scale.
- Use $a = 0$, $? = 0$, and $c = 2$ for all items in scales Q1 and G.
- Use $a = 0$, $? = 2$, and $c = 2$ for all items in scales L and O; Item 7 of the N scale; Items 4, 7, and 8 of the Q4 scale; and Item 1 of the A scale.

Limitations

In terms of study limitations, although our results signal a need to empirically assess category ordering, it is unclear how our results with the 16PF5 would generalize to most other personality inventories that do not explicitly discourage the selection of the middle option. This instruction might have exaggerated any tendency for the categories or thresholds to be in the incorrect order. Further, although the model used in this study imposes relatively few restrictions on the data, it is inevitable that some standard assumptions are made, thus opening the possibility of model misspecifications. For example, the true form of relation between item and latent construct might be better represented by some other function, or the items of a putatively unidimensional scale might be better represented by multiple dimensions. Based on previous research with the 16PF5, we judged that any misspecification would be sufficiently minor to justify the use of the models. Nonetheless, it is possible that future theoretical and empirical work will yield psychometric models for the 16PF5 that improve on the models suggested in this study.

One potentially useful future direction would be to explore the use of the middle response category in the context of unfolding models (Chernyshenko, Stark, Chan, et al. 2001; Chernyshenko, Stark, Drasgow, & Roberts, 2007; Drasgow, Chernyshenko, & Stark, 2010; Stark, Chernyshenko, Drasgow, & Williams, 2006). The majority of item response theory modeling in personality to date—including this study—has used a dominance model to represent the relation between the latent trait and item responses (Weekers & Meijer, 2008). Here, it is assumed that the higher a person's trait level, the more likely it is that he or she will endorse a higher response category on an item. In contrast, unfolding models assume an ideal point process in which the probability of endorsing an item or response category increases with proximity of trait levels to the response category location on the latent trait continuum. Empirically, comparisons between ideal point and dominance models to personality data have provided mixed results, with some studies showing support for ideal point models (e.g., Stark et al., 2006, using the 16PF), and others suggesting fit is not ubiquitously improved by fitting ideal point models (Weekers & Meijer, 2008).

Here, although we acknowledge this interesting and ongoing debate regarding whether item responding is best conceptualized as following an ideal point process versus a dominance process (see, e.g., the responses to Drasgow et al., 2010), we choose to focus on the performance of the middle category through the application of traditional dominance models for a number of reasons. First, although it might be beneficial to develop new inventories based on an ideal point process, the 16PF5 was developed under an implicit dominance model. As a result, the vast majority of its items are extreme-worded items that are not conducive to ideal point modeling, which requires items to be located at both extreme and intermediate levels (Brown & Maydeu-Olivares, 2010). For extreme items, there is little practical difference between applying a dominance versus an ideal point model.

Second, it has been argued that by virtue of including a middle response option, the tenets of ideal point processes are violated. This was noted by Dalal, Carter, and Lake (2014), who argued that the ideal point response process requires individuals to agree or disagree based on the proximity of an item on the trait continuum to their own position. Thus, a *disagree* response could either be because an item is too low or too high on a measured trait. Given this basic principle, Dalal et al. argued it is difficult to conceptualize what a *neither agree nor disagree* statement would mean for the location of an individual's ideal point. They further noted that the middle response category was introduced to Likert-type scales to convey a neutral standing on a trait. Within an ideal point modeling, a neutral standing is conveyed by agreement to items that are at the center of the trait continuum—thus a neutral position is determined by items, not response scales. To demonstrate their concerns with middle response options in ideal point models, Dalal et al. (2014) used a quasi-experimental design where participants responded to a scale developed from an ideal point perspective that had either a 4-, 5-, or 6-point response scale. It was found that whereas the 4- and 6-point scales provided good fit to an ideal point model, the 5-point scale did not.

References

Aluja, A., & Blanch, A. (2004). Replicability of first-order 16PF-5 factors: An analysis of three parcelling methods. *Personality and Individual Differences*, *37*, 667–677.

Aluja, A., Blanch, A., & García, L. F. (2005). Reanalyzing the 16PF-5 second order structure: Exploratory versus confirmatory factorial analysis. *European Journal of Psychology of Education*, *20*, 343–353.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29–51.

Booth, T., & Irwing, P. (2011). Sex differences in the 16PF5, test of measurement invariance and mean differences in the US standardization sample. *Personality and Individual Differences*, *50*, 553–558.

Brown, A., & Maydeu-Olivares, A. (2010). Issues that should not be overlooked in dominance versus ideal point controversy. *Industrial and Organizational Psychology*, *3*, 489–493.

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*, 1–29.

Chernyshenko, O. S., Stark, S., & Chan, K. Y. (2001). Investigating the hierarchical factor structure of the fifth edition of the 16PF: An application of the Schmid–Leiman orthogonalization procedure. *Educational and Psychological Measurement*, *61*, 290–302.

Chernyshenko, O. S., Stark, S., Chan, K. Y., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research*, *36*, 523–562.

Chernyshenko, O. S., Stark, S., Drasgow, F., & Roberts, B. W. (2007). Constructing personality scales under the assumptions of an ideal point response process: Toward increasing the flexibility of personality measures. *Psychological Assessment*, *19*, 88–106.

Conn, S. R., & Rieke, M. L. (1994). *The 16PF fifth edition technical manual.* Champaign, IL: Institute for Personality and Ability Testing.

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual.* Odessa, FL: Psychological Assessment Resources.

Dalal, D. K., Carter, N. T., & Lake, C. J. (2014). Middle response scale options are inappropriate for ideal point scales. *Journal of Business and Psychology*, *29*, 463–478.

Dancer, L. J., & Woods, S. A. (2006). Higher-order factor structures and intercorrelations of the 16PF5 and FIRO-B. *International Journal of Selection and Assessment*, *14*, 385–391.

Drasgow, F., Chernyshenko, O. S., & Stark, S. (2010). 75 years after Likert: Thurstone was right! *Industrial and Organizational Psychology*, *3*, 465–476.

DuBois, B., & Burns, J. A. (1975). An analysis of the meaning of the question mark response category in attitude scales. *Educational and Psychological Measurement*, *35*, 869–884.

Ellis, B. B., & Mead, A. D. (2000). Assessment of the measurement equivalence of a Spanish translation of the 16PF questionnaire. *Educational and Psychological Measurement*, *60*, 787–807.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Glas, C. A. (2010). *Preliminary manual of the software program Multidimensional Item Response Theory (MIRT)*. Enschede, The Netherlands: University of Twente.

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe* (Vol. 7, pp. 7−28). Tilburg, The Netherlands: Tilburg University Press.

González-Romá, V., & Espejo, B. (2003). Testing the middle response categories "Not sure," "In between" and "?" in polytomous items. *Psicothema*, *15*, 278−284.

Hernández, A., Drasgow, F., & González-Romá, V. (2004). Investigating the functioning of the middle category by means of a mixed-measurement model. *Journal of Applied Psychology*, *89*, 687−699.

Hernández, A., Espejo, B., González-Romá, V., & Gómez-Benito, J. (2001). Escalas de respuesta tipo Likert: ¿es relevante la alternativa indiferente? [Likert-type response scales: Is the response category indifferent relevant?]. *Metodología de Encuestas*, *2*, 135−150.

Irwing, P., Booth, T., & Batey, M. (2014). An investigation of the factor structure of the 16PF, Version 5: A confirmatory factor and invariance analysis. *Journal of Individual Differences*, *35*, 38−46.

Kulas, J. T., & Stachowski, A. A. (2009). Middle category endorsement in odd-numbered Likert response scales: Associated item characteristics, cognitive demands, and preferred meanings. *Journal of Research in Personality*, *43*, 489−493.

Kulas, J. T., & Stachowski, A. A. (2013). Respondent rationale for *neither agreeing nor disagreeing*: Person and item contributors to middle category endorsement intent on Likert personality indicators. *Journal of Research in Personality*, *47*, 254−262.

Kulas, J. T., Stachowski, A. A., & Haynes, B. A. (2008). Middle response functioning in Likert-responses to personality items. *Journal of Business and Psychology*, *22*, 251−259.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149−174.

McFadden, L. S., & Krug, S. E. (1984). Psychometric function of the "neutral" response option in clinical personality scales. *Multivariate Experimental Clinical Research*, *7*, 25−33.

Mellenbergh, G. J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement*, *19*, 91−100.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159−176.

Preston, K., & Reise, P. (2014). Detecting faulty within-item category functioning within the nominal response model. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 386−405). New York, NY: Routledge/Taylor & Francis Group.

Preston, K., Reise, S., Cai, L., & Hays, R. D. (2011). Using the nominal response model to evaluate response category discrimination in the PROMIS emotional distress item pools. *Educational and Psychological Measurement*, *71*, 523−550.

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

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111−164.

Rammstedt, B., & Krebs, D. (2007). Does response scale format affect the answering of personality scales? Assessing the Big Five dimensions of personality with different response scales in a dependent sample. *European Journal of Psychological Assessment*, *23*, 32−38.

Reif, M. (2014). *mcIRT: IRT models for multiple choice items. R package version 0.40*. Retrieved from https://github.com/manuelreif/mcIRT

Rossier, J., Meyer de Stadelhofen, F., & Berthoud, S. (2004). The hierarchical structures of the NEO PI−R and the 16PF5. *European Journal of Psychological Assessment*, *20*, 27−38.

Samejima, F. (1972). A general model for free-response data. *Psychometrika Monograph*, *18*.

Samejima, F. (1996). Evaluation of mathematical models for ordered polychotomous responses. *Behaviormetrika*, *23*, 17−35.

Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. A. (2006). Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology*, *91*, 25−39.

Thissen, D., Steinberg, L., & Fitzpatrick, A. R. (1989). Multiple-choice models: The distractors are also part of the item. *Journal of Educational Measurement*, *26*, 161−176.

Weekers, A. M., & Meijer, R. R. (2008). Scaling response processes on personality items using unfolding and dominance models. *European Journal of Psychological Assessment*, *24*, 65−77.