J Bus Psychol (2008) 22:251–259
DOI 10.1007/s10869-008-9064-2

Middle Response Functioning in Likert-responses to Personality Items

John T. Kulas · Alicia A. Stachowski · Brad A. Haynes

Published online: 24 January 2008

Abstract Two studies examined whether the middle response option in graphic rating scales indicates a moderate standing on a trait/item, or rather a "dumping ground" for unsure or non-applicable (N/A) responses. Study One identified middle response-option dysfunction. Study Two indicated that respondents use the middle response option as an N/A proxy, even under implicit 'skip if you do not know' instructional sets. Although middle response category 'misuse' did not adversely affect reliability and validity in these studies, it is recommended that assessment developers (especially in on-line administration contexts) regularly include an N/A response option when administering graphic rating scales.

Keywords Graphic rating Scale · Likert-type · Personality · Assessment · Rasch

Introduction

The typical graphic rating scale used in psychological assessment is comprised of between five and seven response options that indicate a level of agreement or disagreement with an item prompt. The scoring of this type of scale seems straight forward—numbers are assigned consecutively to each progressive category (cf., DuBois and Burns 1975; Hofacker 1984). A total or average is then computed and scales constructed. With an odd-numbered graphic rating scale, mid-point semantic anchors typically indicate neutrality ('neutral') or ambivalence ('neither agree nor disagree').

The current paper investigates whether the middle response option is used to indicate a moderate standing on a trait/item, or rather is viewed by the respondent as a "dumping ground" for unsure or non-applicable response. This would occur, for instance, if the respondent did not view the middle response option as existing along the agreement continuum. If the middle response option is used as a "dumping ground", then the propriety of scoring this option with the consecutive integer protocol is questionable.

J. T. Kulas (✉)
Department of Psychology, St. Cloud State University, Whitney 303, St. Cloud, MN 56301, USA
e-mail: jtkulas@stcloudstate.edu

A. A. Stachowski
George Mason University, Fairfax, VA, USA

B. A. Haynes
Meyecon, Atlanta, GA, USA

Graphic Rating Scales in Psychological Assessment

Graphic rating scales have many variations, with the meaning and absolute number of semantic anchors differing based on researcher preference. Preston and Colman (2000) investigated the reliability, validity, and discriminating power of the *number* of response categories. The authors concluded that the optimal number of options depends on the purpose of the instrument. Scales with 7–10 response options had the highest reliability estimates.

The meaning of the middle point on these odd-optioned scales, the 'neutral' or 'neither' category, is not well understood. Researchers have posited that this middle category is possibly interpreted by respondents in several different ways, some obviously differently than intended (Hofacker 1984). Shaw and Wright (1967) originally

posited three orientations researchers may take in the interpretation of a middle category endorsement: (1) participants having no attitude regarding the attitude object, (2) participants are 'balanced' in terms of evaluation of the attitude object, and (3) the participant's attitude is not clearly defined.

> Stone (2004) sees a problem in the use of the middle point as well:
> Fascination for the popular five-point rating scale 1–2–3–4–5 seems based more on numerology than reason. A middle response choice of "3" can reflect a decision not to prefer either end, a lack of information by which to choose, or an unwillingness to commit to a definitive response. Which is it? How can we understand what "3" means when it can indicate a variety of intentions? I see no value for a middle category given such confusion (Stone, 2004, p. 212).

DuBois and Burns (1975) make a similar argument in stating that respondents may elect to use this option because of ambivalence (unable to decide whether to agree or disagree, whether the statement has more positive or negative characteristics) or indifference (simply not caring either way). A third possibility is that individuals may also use the neutral or middle category because they "do not feel competent enough or sufficiently informed to take a position" (DuBois and Burns 1975).

The job descriptive index (JDI) is an assessment that is used in organizational contexts. The JDI assesses job satisfaction and has three rating categories that were very early rejected as being represented by an interval-scale analog (cf., Hanisch 1992; Smith et al. 1969). The JDI scores the neutral ('?' in this case) category as being more negative than positive (i.e., satisfaction scores a '3', lack of satisfaction scores a '0', and a response of '?' is scored as a '1'). This approach was taken after it was discovered that dissatisfied individuals tended to use the '?' more often than did satisfied individuals. An interval approach would score the '?' as a '1.5' rather than a '1'.

Asking a similar question, McFadden and Krug (1984) investigated the clinical analysis questionnaire (CAQ) and found that psychotic individuals tend to endorse the middle response option almost twice as frequently as do non-psychotics; Neurotic individuals and those suffering from personality disorders also choose to endorse the neutral category more frequently than "normals", but not as frequently as do psychotic individuals. The possibility of a correlation of neutral category endorsement with the trait/construct being measured is certainly problematic from a traditional scoring perspective and suggests that the typical graded rating scale scoring protocol is not entirely appropriate.

Summary and Hypotheses

Two investigations were completed to answer questions surrounding the functioning of middle response options. In Study One, an archival data set was investigated for functioning peculiarities of the middle response option category. For Study Two, a test-retest procedure was implemented, with differences across assessment administrations consisting of the presence or absence of a 'not applicable' (N/A) response option.

Hypothesis 1 (Study One) Threshold disorderings will be found because of suboptimal use of the middle response option category.

Consistent with past research (Hofacker 1984; Schriesheim and Schriesheim 1974), the graphic rating scale is expected to be ordinal (using Stevens' (1946) taxonomy). Additionally, the functioning of the middle response option category is expected to be shown to be problematic. Suboptimal response frequency of the middle response option category is expected to manifest in disordinal threshold orderings. Study two explores whether or not this threshold violation, if found, is problematic.

Hypothesis 2 (Study Two) Individuals who endorse a N/A response option will have a proclivity toward choosing the middle response scale category when not offered the N/A alternative.

It is thought that the middle response option category on typical odd-numbered graphic rating scale formats is (at least occasionally) used as a "dumping ground" for unsure or non-applicable responses. The theory here is that the commonly administered five-point Likert scale does not always elicit a progressively ordered continuum of response. The semantics 'neither' or 'neutral' don't necessarily indicate halfway between agreement and disagreement. In order to test this hypothesis, N/A item endorsements will be obtained, and a second administration of the item will be given without offering the N/A option.

Hypothesis 3a (Study Two) Distributions of scale-level ability estimates will exhibit less variance when the N/A response option is not made available.

If respondents are likely to choose a neutral response option category when unsure of their true standing on a trait, then the distribution of these respondents' aggregated scale scores could exhibit attenuated variability estimates. It is expected that if 'neutral' is used as a dumping ground, more '3''s will be used which means less extreme endorsements are used, which means variance estimates will be smaller in magnitude. This effect is expected to be small, as the current assessment has only 20 items per scale, and the would-be use of the N/A response-option

category is not expected to occur with great frequency. A null finding of this hypothesis would be informative as well, as a lack of differentiation in scale-level variability estimates would indicate that the absence of a N/A option is not problematic (with regard to scale-level variability). Support of Hypothesis 3a would lead to potential attenuation of criterion-related validity coefficients, therefore, our final hypothesis:

Hypothesis 3b (Study Two) Scale validities will be attenuated when coefficients are computed using scale scores that do not offer the N/A response option.

Study One

Method

Participants

Twenty one thousand five hundred and eighty-eight individuals completed a 300-item web-based personality assessment between August, 1999 and March, 2000. Seven thousand eight hundred and fifty-nine (36.4%) were male and 13,729 (63.6%) were female. The mean age was 26.24 ($SD = 10.79$).

Materials

The international personality item pool (IPIP) is an internet-housed item bank which at the time of investigation consisted of 2,036 indicators of 280 personality 'scales'. The items and scales of the IPIP are available to all (they are free of copyright restrictions) at http://ipip.ori.org/. All Study One respondents provided internet-based responses to 300 items. These 300 items represent an IPIP version of Costa and McCrae's NEO-PI-R (1992). This 300 IPIP-item Big 5 structure was first identified by Goldberg (1999)—for this study an on-line version was used (cf., Johnson 2005). The response format consisted of a five-point graphic rating scale: 1 (very inaccurate), 2 (moderately inaccurate), 3 (neither accurate nor inaccurate), 4 (moderately accurate), and 5 (very accurate).

Procedure

Potential problems in the nature of functioning of the 'middle' response category were investigated using thirty 10-item sub-scales of the "IPIP-NEO" (Johnson 2005, p. 113). This was accomplished by estimating probabilities of response option category endorsement through an application of Rasch modeling technology.

The General Rasch Model. For an assessment comprised of dichotomously scored (e.g., correct/incorrect) items, the Rasch model can be represented by:

$$\log\left(\frac{P_{ni1}}{P_{ni0}}\right) \equiv B_n - D_i$$

where B_n is the ability level (B) of person n, D represents an item's (i) difficulty level, P_{ni1} gives the probability of person n responding correctly (1) to item i, and P_{ni0} is the probability of the same person responding incorrectly (0). The model therefore specifies the nature of relationship between two parameters: (1) item difficulty, and (2) person ability, and the probability of a correct response to an item by a given person. The specific metric of the probabilities is referred to as a log odds (called logits) scale. Item calibration takes place through identifying each item's location along a latent dimension being assessed (here the latent dimensions are personality traits), as well as the item's location with respect to other items. Both person and item parameters are estimated on the common logit scale.

The rating scale model (Andrich 1978) is an extension of the general Rasch model referred to above and is specified as:

$$\log\left(\frac{P_{nik}}{P_{ni(k-1)}}\right)/B_n - D_i - F_k$$

In this application, the specification of a correct (1) or incorrect (0) response is replaced with a response in category k or $k-1$. The F term represents a threshold parameter corresponding to a location where the 'k' and '$k-1$' category endorsements are equal.

The rating scale model, applied to the current data, allows for an investigation of rating scale functioning. For the purposes of Study One, the aspect of rating scale functioning of primary interest was the probability of utilizing the middle response category, relative to the probabilities associated with using the adjacent 'moderately accurate' and 'moderately inaccurate' categories.

Results

Figure 1 shows a Rasch estimated category probability curve for one 10-item scale. The x-axis represents standing along the latent trait of interest (here Anxiety; defined by the 10 items presented in Table 1). The y-axis represents the probability of responding to any category (ranging from 0 [not likely] to 1 [extremely likely]). The probability of, for example, 'Very Inaccurate' endorsement, decreases as the trait approaches higher levels of anxiety. The middle response options have the highest probability of endorsement at moderate levels of anxiety, with lower probabilities of endorsement as respondents become both more and less anxious.

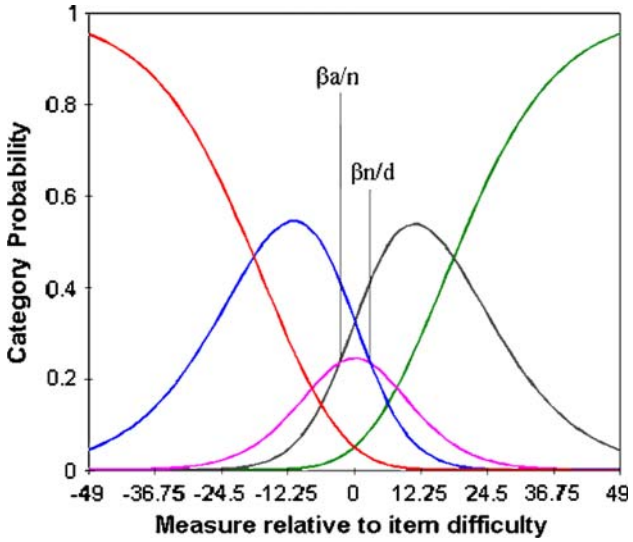

Fig. 1 Category probability curves for the 5 response option categories for IPIP 'anxiety' items ($k = 10$, $N = 21{,}588$)

The functions of note are these middle functions. There is a disordering of response *thresholds*, such that the threshold between the 'Moderately Inaccurate' and 'Neither Inaccurate nor Accurate' response categories ($\beta_{n/d}$) lies above the threshold between the 'Neither Inaccurate nor Accurate' and 'Moderately Accurate' response categories ($\beta_{a/n}$). This means that the empirical category thresholds are 'reversed' from their intended/predicted/logical ordering (Andrich 2004). Figure 1 shows that this disordinal empirical ordering is caused by a suboptimal 'neither' response category.

Even an individual whose anxiety level is at the trait mid-point (zero on the *x*-axis) has a higher probability of endorsing the 'moderately accurate' or 'moderately inaccurate' response option categories than of endorsing the 'neither' response option. Why include 'neither' as an option at all if its probability of endorsement, *even at the middle of the trait*, is lower than the probability of endorsing either of the polarized options. Nineteen of 30 subscales (63.3%) exhibited this disordinal response scale threshold ordering, such that $\beta_{a/n}$ did not lie 'to the right' of $\beta_{n/d}$. All investigated subscales are presented in Table 2. Implications are addressed in the discussion. Study Two investigates one possible reason for middle/neutral/neither response option 'dysfunction'—the use of the middle category option as a 'dumping ground' for unsure or not applicable responses.

Study Two

Participants

One hundred thirty undergraduate students at a large Midwestern University participated for class extra-credit.

Materials

A suggested 100-item personality measure was selected from the website of the IPIP. The personality measure contains 20 items per Big 5 personality dimension. In order to examine the degree to which item complexity contributes to the use of the middle response category as a 'dumping ground', five items per personality dimension (25 total items) were modified through replacing item verbs with less common synonyms. The original and modified content of the 25 altered items is presented in Table 3. The response scale used was the more familiar: 1 (strongly disagree), 2 (disagree), 3 (neither agree nor disagree), 4 (agree), and 5 (strongly agree). In addition to these five common response option categories, one form had an

Table 1 Anxiety items included in analysis of category functioning and their empirical response frequencies (rows do not equal 100%, as missing responses have been excluded)

Item	Very inaccurate (%)	Moderately inaccurate (%)	Neither (%)	Moderately accurate (%)	Very accurate (%)
1. Worry about things	3.8	12.8	13.3	**44.7**	25.3
2. Fear for the worst	13.3	24.2	19.4	**27.8**	14.6
3. Am afraid of many things	23.4	**33.2**	19.6	17.9	5.8
4. Get stressed out easily	12.7	23.5	15.9	**28.4**	19.2
5. Get caught up in my problems	5.2	18.7	17.9	**41.1**	16.8
6. Am not easily bothered by things	10.5	26.5	15.1	**31.8**	15.6
7. Am relaxed most of the time	11.6	**43.8**	16.9	21.3	5.9
8. Am not easily disturbed by events	7.5	31.1	20.6	**32.2**	8.1
9. Don't worry about things that have already happened	9.3	26.1	10.5	**36.3**	17.5
10. Adapt easily to new situations	20.5	**48.4**	13.7	13.5	3.5

Table 2 IPIP subscales included in analysis of category functioning (bold indicates threshold disordering)

10-Item subscale	Scale
Anger	**Neuroticism**
Anxiety	**Neuroticism**
Depression	Neuroticism
Immoderation	**Neuroticism**
Self-consciousness	**Neuroticism**
Vulnerability	**Neuroticism**
Activity level	**Extraversion**
Assertiveness	**Extraversion**
Cheerfulness	Extraversion
Excitement-seeking	**Extraversion**
Friendliness	**Extraversion**
Gregariousness-	Extraversion
Adventurousness	**Openness to experience**
Artistic interests	**Openness to experience**
Emotionality	**Openness to experience**
Imagination	**Openness to experience**
Intellect	Openness to experience
Liberalism	Openness to experience
Altruism	Agreeableness
Cooperation	**Agreeableness**
Modesty	**Agreeableness**
Morality	Agreeableness
Sympathy	Agreeableness
Trust	Agreeableness
Achievement striving	Conscientiousness
Cautiousness	**Conscientiousness**
Dutifulness	**Conscientiousness**
Orderliness	**Conscientiousness**
Self-discipline	**Conscientiousness**
Self-efficacy	Conscientiousness

additional category of: N/A (not sure/not applicable). These two forms are heretofore referred to as 'traditional' (the 5-response option form) and N/A (the traditional form with the addition of the N/A option). A third questionnaire, a criterion instrument, asked participants to rate themselves on five personality dimensions on a scale of 1 (indicated having more of the trait) to 10 (indicated having little of the trait). The assessments were administered on-line.

Procedure

Participants were scheduled for administrations 1 week apart. The presentation of the N/A versus traditional form was counterbalanced across participants. After the second administration, participants completed the 5-item criterion questionnaire.

Results

Twelve participants did not respond to both time 1 and time 2 administrations, yielding a final test-retest sample of 118 individuals. Although the N/A option endorsement was infrequent (1%) in relation to the absolute number of responses given, 32 of the 118 valid respondents (27%) endorsed the N/A option category at least once. Table 4 presents the 25 items (out of 100 administered; 25%) that elicited N/A responses. Nineteen of the 25 (76%) were reworded items.

Figure 2 shows that there was an overwhelming tendency to choose the middle response option category on the traditional form if N/A was chosen for the same item on the N/A form ($\chi^2_{(4)} = 252.27$, $p < .05$). Note that individuals tended to choose this middle option overwhelmingly ($n = 108$) in favor of simply 'skipping' the item response ($n = 8$). Hypothesis 2 was supported. These results have potential implications for the validity of measures.

Scale Variance Attenuation

In order to assess the effect of attenuation of variance for scale-level scores (Hypothesis 3a), three sets of analyses were performed using two different scoring protocols. In the first set of three analyses, a middle response option was scored as a '3' and N/A's were treated as missing values. In the second set of analyses, both middle response options and N/A's were treated as missing values. Because similar effects were recorded across both sets of analyses, only the first set (using the '3' scoring protocol) is reported here.

Standard Deviation Scores for N/A Versus Traditional Endorsers. Scale-level standard deviations were computed for all respondents (for both forms). A *t*-test was conducted for those who chose N/A at least once versus those who did not choose N/A. None of the 10 *t*-tests looking at mean differences in *SD* scores approached the liberal .05 level of significance. *T*-values ranged from a low of .02 (openness to experience scale; form 2) to a high of 1.51 (agreeableness scale; form 2).

Correlations of Endorsing N/A and Subsequent Scale SD. Correlations were computed between number of N/A's chosen and individual standard deviation scores. This was once again done across each of the 10 administered scales. None of these correlations reached significance (absolute values of coefficients ranged from .00 to .14 [*n*'s ranged from 117 to 118]).

Table 3 Altered items included in the IPIP assessment (five altered items per Big 5 dimension)

Big 5 dimension	Altered (included) IPIP item	Original IPIP item
Extraversion	21. ***Initiate*** conversations	21. Start conversations
	26. Have little to ***utter***	26. Have little to say
	76. ***Suppress*** my feelings	76. Bottle up my feelings
	86. Am a very ***clandestine*** person	86. Am a very private person
	96. Am ***proficient*** in handling social situations	96. Am skilled in handling social situations
Agreeableness	2. ***Affront*** people	2. Insult people
	12. Am not interested in other people's ***quandaries***	12. Am not interested in other people's problems
	17. ***Commiserate*** with others' feelings	17. Sympathize with others' feelings
	52. Am ***apathetic*** to the feelings of others	52. Am indifferent to the feelings of others
	67. Know how to ***placate*** others	67. Know how to comfort others
Conscientiousness	13. Pay attention to ***minutiae***	13. Pay attention to details
	18. Make a ***shambles*** of things	18. Make a mess of things
	48. ***Disregard*** my duties	48. Neglect my duties
	58. ***Squander*** my time	58. Waste my time
	78. Find it ***arduous*** to get down to work	78. Find it difficult to get down to work
Neuroticism	4. Get ***frazzled*** easily	4. Get stressed out easily
	9. Am ***unperturbed*** most of the time	9. Am relaxed most of the time
	19. Seldom feel ***melancholy***	19. Seldom feel blue
	29. Am not easily ***perturbed*** by things	29. Am not easily bothered by things
	54. Have ***recurrent*** mood swings	54. Have frequent mood swings
Openness to experience	15. Have a ***flamboyant*** imagination	15. Have a vivid imagination
	10. Have difficulty understanding ***intangible*** ideas	10. Have difficulty understanding abstract ideas
	40. Try to avoid ***byzantine*** people	40. Try to avoid complex people
	90. Am ***proficient*** at many things	90. Am good at many things
	95. Love to read ***exigent*** material	95. Love to read challenging material

SD Values Across Administered Forms. Finally, paired samples *t*-tests were computed for standard deviation scores across forms (traditional versus N/A). This was done twice—once for all respondents ($t_{ex} = .07$; $t_{ag} = -.85$; $t_{co} = 1.10$; $t_{ne} = -.38$; $t_{op} = 1.30$; all p's $> .05$) and once for only the 33 respondents who chose the N/A response option category at least once ($t_{ex} = .78$; $t_{ag} = -.35$; $t_{co} = -.43$; $t_{ne} = -1.80$; $t_{op} = .92$; all p's $> .05$). Across these three perspectives, scale-level standard deviations are not seemingly affected by the simple presence or absence of an N/A response alternative. Hypothesis 3a was not supported.

Scale Reliability

Test–retest estimates across the two administered forms were .95 (extraversion), .89 (agreeableness), .88 (conscientiousness), .91 (neuroticism), and .90 (openness to experience). Internal consistency estimates for the traditional and N/A scales were comparable: extraversion ($\alpha_{N/A} = .93$, $\alpha_{trad} = .93$), agreeableness ($\alpha_{N/A} = .75$, $\alpha_{trad} = .77$), conscientiousness ($\alpha_{N/A} = .90$, $\alpha_{trad} = .89$), neuroticism ($\alpha_{N/A} = .93$, $\alpha_{trad} = .94$), and openness to experience ($\alpha_{N/A} = .85$, $\alpha_{trad} = .87$).

Validities. Correlations were computed between scale scores and the corresponding single item, 10-response point self-indicator criteria for both traditional and N/A forms. Differences between validity coefficients (across forms) were computed using Steiger's (1980) formula for nonindependent correlation coefficients:

$$t = (r_{12} - r_{13}) \sqrt{\frac{(N-1)(1+r_{23})}{2\frac{(N-1)}{(N-3)}|R| + \frac{(r_{12}+r_{13})^2}{4}(1-r_{23})^3}}$$

This analysis was performed on all five scales using all 118 respondents, as well as using only the 32 respondents who chose the N/A response option at least once. Results are reported in Table 5. As can be seen through inspection of the table, nine of the 10 validity coefficient differences were in the opposite direction as predicted. None of these coefficients exceeded significance criteria after implementing a Bonferonni correction for 10 analyses ($\alpha = .005$). Three of these coefficient differences exceeded significance criteria using a liberal alpha of .05. Hypothesis 3b was not supported.

Table 4 Items eliciting at least one N/A response

Item	Number of N/A responses
86. Am a very clandestine person	24
40. Try to avoid byzantine people	19
67. Know how to placate others	13
95. Love to read exigent material	11
13. Pay attention to minutiae	10
78. Find it arduous to get down to work	9
38. Shirk my duties	7
17. Commiserate with others' feelings	6
9. Am unperturbed most of the time	5
29. Am not easily perturbed by things	5
58. Squander my time	5
19. Seldom feel melancholy	4
2. Affront people	3
12. Am not interested in other people's quandaries	3
26. Have little to utter	3
52. Am apathetic to the feelings of others	2
53. Am exacting in my work	2
4. Get frazzled easily	1
10. Have difficulty understanding intangible ideas	1
15. Have a flamboyant imagination	1
16. Keep in the background	1
57. Make people feel at ease	1
90. Am proficient at many things	1
93. Love order and regularity	1
99. Grumble about things	1

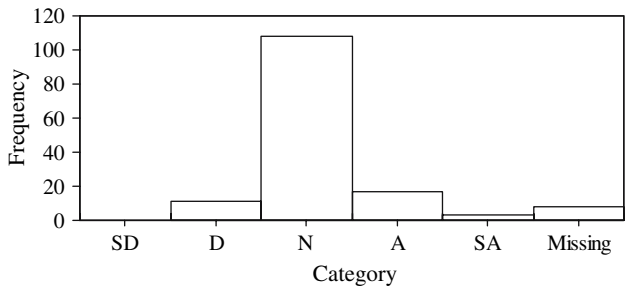

Fig. 2 Traditional form responses (when N/A form response equaled N/A)

Discussion

Initially, graphic rating scales, as specified by Likert (cf., Likert 1932), had two different scoring options. One procedure was based on assigning consecutive integers, the other took empirical response frequencies into consideration. The evolution of the scoring protocol has been that assessment developers follow the consecutive integer approach. The consecutive integer method for scoring

Table 5 Differences between validity coefficients for traditional and N/A scale scores

Dimension	Validity coefficient (N/A form)	Validity coefficient (traditional form)	Test–retest coefficient	*t*
Openness (all)	.40	.43	.90	−.70
(N/A respondents)	.45	.52	.89	−.90
Consc. (all)	.57	.64	.88	**−1.84**
(N/A respondents)	.71	.78	.90	−1.29
Extraversion (all)	.78	.79	.95	−.37
(N/A respondents)	.73	.77	.93	−.99
Agree. (all)	.19	.18	.89	.23
(N/A respondents)	.35	.35	.89	−.11
Neuroticism (all)	.45	.52	.91	**−1.87**
(N/A respondents)	.42	.56	.92	**−2.40**

Significant *t*'s are bolded (.05 one-tailed)

responses is suspect, however, if the response scale midpoint is not viewed by respondents as representing a simple midpoint along a continuum of agreement.

When implementing psychometric models of rating scale functioning, it is not the category itself that is of primary interest, but the location of the threshold between categories. The question involves the transitional boundary between, for instance, 'very inaccurate' and 'moderately accurate'. With five response option categories, four transitional boundaries exist (the two extreme response categories have no upper/lower bound). In the current study, over 60% of investigated scales yielded peculiar response threshold orderings. The boundaries between the most extreme and next extreme categories were located as expected, but the thresholds between the moderate and neutral categories were 'reversed'.

This does not mean that the threshold disordering is problematic. It may merely be an artifact of the empirical response frequencies. The middle response option is simply less likely (at all anxiety levels) than are other response options. For each of the 10 items defining the anxiety scale (for example), an inspection of individual item response frequencies indicates that other options were empirically preferred to the neutral response option category. Table 1 shows that the middle response option (for the demonstrated anxiety scale items) was always at most the 3rd most frequently chosen response category. For 3 of the 10 investigated items, the middle response option was the 4th most frequent choice.

Taken with study two's validity results, this seems to be a problem more for accurate psychometric models of response (category threshold parameter identification) than it is for psychometric issues of reliability and validity. Nevertheless, middle category endorsement from this perspective is *potentially* problematic, as it may result from a multitude of respondent motivations (i.e., what does endorsement *mean*?).

Study two demonstrates that the neutral category is at least sometimes used as a 'dumping ground' for unsure or non-applicable responses, as there was a tendency to choose the *neutral* response option category on the traditional form if N/A was indicated on the N/A form. This effect should be especially prevalent in situations in which the reading level of the test takers is 'low' compared to the reading level of the assessment, as the majority of neutral/not applicable cross endorsements came on items whose content had been edited by replacing prompt verbs with less common synonyms.

Summary and Implications

One practical implication to be taken from this simple study involves the choice to include or exclude the middle option when constructing scales. Indices of variability did not differ based on inclusion or exclusion of the 'middle' response category in scale-level scoring. Additionally, correlations of scale scores computed without scoring the middle response option were nearly identical to scale scores computed with inclusion of the '3' option. Correlations ranged from .94 to 1.0 (even though fewer items contributed to scale scores when the '3' option was coded as missing). This suggests one option available to researchers who are concerned with the ambiguity associated with middle response-option endorsement: exclude these moderate scores from scale-construction. This could either be done by exclusion of the '3' category in scoring, or by using a '−2', '−1', '1', '2' scoring protocol. The use of a $1 \rightarrow 5$ rather than a $-2 \rightarrow +2$ analog is likely preferable in feedback contexts, where negative connotations could be inferred from a negatively valenced scale score. A second practical implication involves the choice to include or not include an N/A option for respondents. This option would seem preferable to the no- or alternative-scoring options, as a higher number of scale-items would be retained for scale identification.

It is possible that, with Likert-type indicators of personality constructs, there is a confounding of the construct(s) being measured with the process used to measure the construct(s). That is, the Likert-type response scale may not be entirely orthogonal to the construct(s) being measured. Is it not possible, in the current context, that a neurotic individual would be more likely to choose the 'neither' category than would a non-neurotic individual (or conversely, a conscientious person would avoid the use of the uncertain category?).[1] Test developers should carefully consider response scales and their potential relationship with the construct(s) being measured, and given this possibility, an "N/A" option should surely be presented to respondents on personality assessments.

Although respondents will sometimes use the middle response option as an N/A proxy, this misuse did not adversely affect reliability and validity estimates in the current study. The use of the middle category as a dumping ground should be viewed as problematic beyond the ambiguity involved with the *meaning* of the response, however, as the lack of effect on reliability and validity coefficients in the current study could have been attributable to the low overall frequency with which respondents chose the N/A response option category. One situation in which respondents would be expected to choose N/A was manipulated in the current study: item complexity. Another environment in which the N/A/middle category confound is expected to be prominent is in on-line assessment.

As more instruments move toward on-line administration formats, more respondents become test-savvy with regards to implicit expectations of response. That is, as on-line assessments become more prevalent and standardized, expectations of respondents should become more well-formulated as they increase their own personal exposure to such assessments. It is not uncommon to encounter on-line questionnaires that 'bump' respondents back if fields are not completed. The prevalence of this on-line procedure would be expected to create response tendencies in frequent test-takers: respond to all items. The instructional set of 'respond to the best of your ability' used in the current study carries an implicit message of 'do not respond if you are unable'. The instructions did not have an effect on either the tendency to skip items or the middle response/N/A relationship. Without the availability of an N/A option and with an implicit (and sometimes explicit) rule to answer all items, scale-level scores are possibly being influenced by 'untrue' middle response category endorsement. It is therefore recommended that item developers (especially in on-line administration contexts) include an N/A option, as this simple inclusion should address one potential area of ambiguity regarding middle-response endorsement.

[1] This is in fact the case looking at the correlations of middle category endorsement (across 300 items) with Big 5 scale scores for Study One's 21,588 participants ($r_{neur} = .11$, $r_{con} = -.20$, $r_{agree} = -.11$, $r_{open} = -.24$, $r_{ext} = -.14$, all p's $< .0001$).

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561–574.

Andrich, D. (2004). Understanding resistance to the data-model relationship in Rasch's paradigm: A reflection for the next generation. In E. V. Smith Jr. & R. M. Smith (Eds.), *Introduction to Rasch Measurement* (pp. 167–200). Maple Grove, MN: JAM Press.

Costa, P. T. Jr., & McCrae, R. R. (1992). Revised NEO personality inventory (NEO PI-R^{TM}) and NEO five-factor inventory (NEO-FFI) professional manual. Odessa, FL: Psychological Assessment Resources.

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

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.

Hanisch, K. A. (1992). The job descriptive index revisited: Questions about the question mark. *Journal of Applied Psychology, 77*, 377–382.

Hofacker, C. F. (1984). Categorical judgment scaling with ordinal assumptions. *Multivariate Behavioral Research, 19*, 91–106.

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from Web-based personality inventories. *Journal of Research in Personality*, *39*, 103–129.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 22*(140), 5–55.

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.

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales; reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica, 104*, 1–15.

Schriesheim, C., & Schriesheim, J. (1974). Development and empirical verification of new response categories to increase the validity of multiple response alternative questionnaires. *Educational & Psychological Measurement, 34*, 877–884.

Shaw, M. E., & Wright, J. M. (1967). *Scales for the Measurement of Attitudes*. New York: McGraw-Hill.

Smith, P. C., Kendall, L. M., & Hulin, C. L. (1969). The measurement of satisfaction in work and retirement: A strategy for the study of attitudes. Skokie, IL: Rand-McNally.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87*, 245–251.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*, 677–680.

Stone, M. H. (2004). Substantive scale construction. In E. V. Smith Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 201–225). Maple Grove, MN: JAM Press.