

Testing the middle response categories «Not sure», «In between» and «?» in polytomous items

Vicente González-Romá and Begoña Espejo
Universidad de Valencia

The first aim of this study is to test two assumptions of integer scoring (the assumption of ordered response categories, and the assumption of ordered thresholds). We tested these assumptions using a set of items extracted from the social boldness scale of the 16PF questionnaire that were presented with three different middle response categories: «Not sure», «In between» and «?». The second aim was to compare a 3-point response scale with a dichotomous response scale in terms of the information function and correlations with external criteria. The sample was composed of 816 undergraduate students. The results obtained showed that both assumptions were met only when the middle response category was «In between». The results also revealed that a 3-point response scale including «In between» provided more information than a dichotomous response scale. Both scales showed similar correlations with the considered external criteria.

Análisis de las categorías de respuesta central «No estoy seguro», «Término medio» y «?» en ítems politómicos. El primer objetivo de este estudio fue poner a prueba dos supuestos del procedimiento de puntuación de ítems politómicos mediante asignación de números enteros: a) el supuesto de categorías de respuesta ordenadas, y b) el supuesto de umbrales ordenados. Se utilizó un conjunto de ítems de la escala de atrevimiento social del test 16PF que fueron presentados con tres categorías de respuesta intermedia diferentes: «No estoy seguro», «Término medio» y «?». El segundo objetivo fue comparar una escala de respuesta politómica de tres alternativas con una dicotómica en términos de información y correlaciones con criterios externos. La muestra se componía de 816 estudiantes universitarios. Los dos supuestos sólo se cumplieron cuando la categoría central fue «Término medio». La escala de tres alternativas de respuesta con «Término medio» como la categoría central, ofreció más información que una escala dicotómica. Las correlaciones de ambas escalas con los criterios externos considerados fueron similares.

Response scales that include a middle category such as «undecided», «not sure» or «?» are often used in Psychology (Cañadas & Sánchez-Bruno, 1998). These middle categories are frequently used to represent a middle position in the latent construct. Subjects respond to this kind of items by selecting the response category that better represents their position with regard to the item content. Successive integers are assigned to successive response categories, so that they reflect increasing levels of the latent construct. This scoring procedure is called *integer scoring* (e.g., Andrich & Schoubroeck, 1989). From this perspective, it is implicitly assumed that a subject who responds «1. Not sure» to an item with a 3-point response scale («0. Disagree», «1. Not sure», «2. Agree») shows a higher level on the measured latent variable than a subject who responds «0. Disagree». The assumptions that underlie this scoring system are: 1. The $(m+1)$ response categories are *ordered* along the continuous latent variable (θ) in the way indicated by the

integers assigned to each response category; 2. There are m *ordered* thresholds (τ) that designate boundary positions between every two adjacent response categories, so that the latent continuum can be viewed as partitioned into $m+1$ intervals by the m thresholds, and; 3. The probability of each subject's response (y) to an item is determined by the interval in which the subject's value in θ falls.

When these assumptions are met, in the case of an item with a 3-point response scale, it is expected that the observed item's score (y) will equal 0 if $\theta < \tau_1$; y will equal 1 if $\tau_1 \leq \theta < \tau_2$, and y will equal 2 if $\theta \geq \tau_2$. Therefore, the score obtained by a subject by means of integer scoring represents the number of ordered thresholds reached by the subject on the continuum of the latent variable (θ) from the origin (Andrich, 1978a,b; Andrich & Schoubroeck, 1989).

The first assumption implies that those subjects with the lowest levels in θ will be the ones with the greatest probability of responding by using the response category with the smallest assigned integer (e. g., «0. Disagree»), those subjects with intermediate levels in θ will be the ones with the greatest probability of responding by using the middle response (e. g., «1. Not sure»), and that those subjects with the highest levels in θ will be the ones with the greatest probability of responding by using the response category with the largest assigned integer (e. g., «2. Agree»). The second assumption (ordered thresholds) implies that every response cate-

gory is the response with the highest probability of being selected within a given interval of θ values.

When both assumptions are met, the category response functions (CRFs) for the case of three response categories and two thresholds show the forms depicted in Figure 1. If the first assumption is not satisfied, then the scoring system is not tenable. Consider the example depicted in Figure 2 in which the assumption of ordered response categories is not satisfied. Those subjects with the lowest levels in θ are not those with the greatest probability of responding by selecting «0. Disagree». Instead, responding by selecting «1. Not sure» is typical for those subjects. In this case, we cannot infer that subjects responding «0. Disagree» have a lower level in θ than those responding «1. Not sure». The order among the response categories suggested in Figure 2 («Not sure», «Disagree», «Agree») is not consistent with the order assumed by the integer scoring.

If the first assumption is satisfied, but the second is not, then the score obtained by a subject cannot represent the number of *ordered* thresholds reached by the subject, as assumed by the integer scoring. Moreover, in this case the response categories are ordered, but one of them is not performing as expected in terms of its probability of being selected, that is, there is one response category that for no interval of θ values is the response with the highest probability of being selected. Consider the example depicted in Figure 3. In this example, the response categories are ordered as assumed by the integer scoring, but the thresholds between them are not. The subjects with the lowest levels in θ are the ones with the greatest probability of responding by selecting «0. Disagree»,

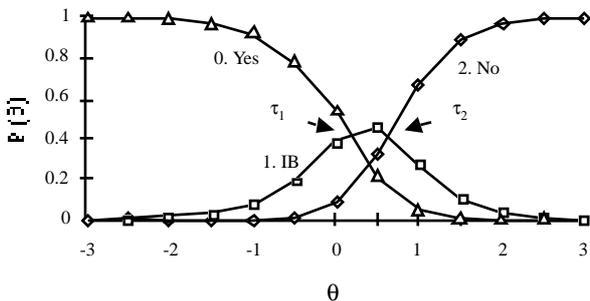


Figure 1. Category response functions for item 107 of scale H_{IB} with ordered response categories and ordered threshold (τ) estimates. Note. IB: In between

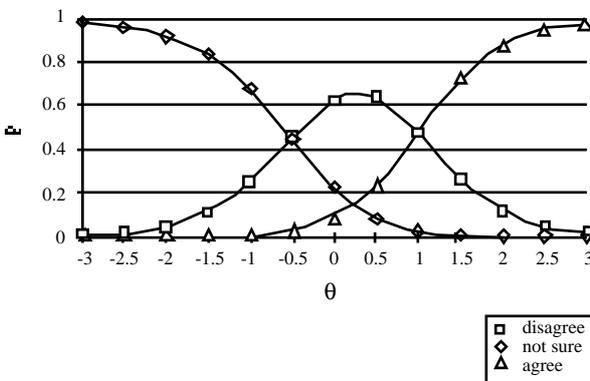


Figure 2. Response functions for an item with three response categories not ordered as assumed by the integer scoring

the subjects with the highest levels in θ are the ones with the greatest probability of responding by selecting «2. Agree», and the subjects with intermediate levels in θ are the ones who have the greatest probability of responding by selecting the middle response («1. Not sure»). However, for the last group of subjects the probability of responding by selecting the middle category is smaller than the probability of responding by selecting other responses.

There is no guarantee that the response categories and the thresholds between them will operate in the way expected from the integer scoring procedure (Andrich et al., 1997). Thus, their assumed ordering must be treated as hypotheses to be tested. Although these hypotheses can be tested by means of different Item Response Theory (IRT) models, this question has seldom been addressed by previous research. Andrich, deJong & Sheridan (1997) tested the hypothesis of ordered thresholds in a response scale with a middle category using the Unidimensional Rasch Model for Ordered Response Categories (URM-ORC) (Andrich, 1978a,b). The attitude of 114 teachers towards a strategy of ‘direct instruction’ teaching was measured by a questionnaire made up of 40 items that were responded to using a 5-point Likert response scale (0. Strongly Disagree (SD), 1. Disagree (D), 2. Not Sure (NS), 3. Agree (A), 4. Strongly Agree (SA)). The attitude was measured before and after teachers were actually exposed to the strategy. The threshold estimates yielded by the URM-ORC revealed that 8 out of 40 items on the pretest, and 19 out of 40 items on the post-test, did not present ordered thresholds. For all these items the disordered thresholds problem involved the middle category (NS), which did not work as a category in the middle of the others. The response functions for this category were similar to those depicted in Figure 3 for the middle category. Therefore, the probability of responding with the middle category was never greater than the probability of responding with the other categories. This means that the NS category was not performing as expected in terms of its probability of being selected. Even people whose location should imply the greatest probability of choosing the middle category had a greater probability of choosing some of the other categories (Andrich et al., 1997). Andrich and colleagues (1997) concluded that all these results «confirm concerns with the middle category designated as Neutral, Not Sure or Undecided in the Likert-style response format, and indicates that in this case it should be not treated as an attitude more or less somewhere between a negative and a positive attitude» (Andrich et al., 1997, p. 66).

Similar results were obtained by Rojas and Fernández (2000) and Hernández, Espejo, González-Romá and Gómez-Benito (2001). Rojas and Fernández (2000) studied whether the assump-

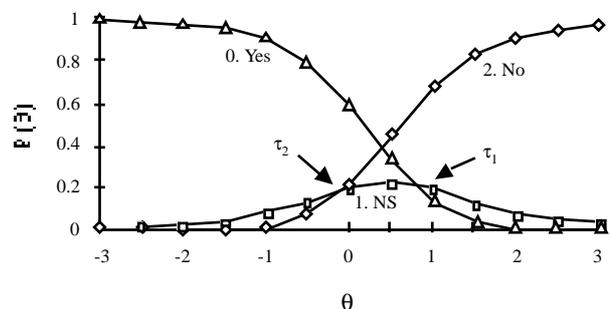


Figure 3. Category response functions for item 107 of scale H_{NS} with ordered response categories and disordered thresholds (τ) estimates. Note. NS: Not sure

tion of ordered thresholds was met when the middle response category was «Neither in agreement nor disagreement». The rating scale model (Wright & Masters, 1982) was fitted to data gathered using a shyness scale composed of 13 items with a 5-point Likert response scale. The results obtained showed that the assumption of ordered thresholds was only supported when subjects assigned to the middle response category the meaning of «an intermediate position in the latent trait», and it was not supported when the assigned meaning was «an expression of doubt and indecision». Hernández and colleagues (2001) tested the assumptions of ordered response categories and ordered thresholds on three short scales of job satisfaction, role overload and support climate that were responded to by using a 5-point Likert scale whose middle response category was «Indifferent». They fitted Bock's (1972) Nominal Model (BNM) to data provided by each scale. The results obtained showed that for all the items within each scale the response categories were ordered. However, the assumption of ordered thresholds was only met in 4 out of the 10 items analyzed.

One of the causes of this disordering of thresholds may be multidimensionality of responses. Some of the factors that may lead respondents to answer by selecting the middle category are ambivalence and indifference (Cronbach, 1946; Dubois & Burns 1975; Edwards, 1946; Goldberg, 1971; Kaplan, 1972; Worthy, 1969). An ambivalent respondent chooses the middle category because he/she has mixed feelings about the object of evaluation and he/she cannot make up his/her mind as to whether he/she agrees or disagrees with the proposed statement. An indifferent respondent chooses the middle category because he/she is minimally involved with the topic of the statement. Other factors that may lead to answering with the middle category are (Cruickshank, 1984; Dubois & Burns, 1975; Worthy, 1969): 1. that subjects show a specific response style, 2) that subjects do not understand the statement, 3) that they do not feel competent enough or sufficiently informed to take a position, 4) that they do not want to reveal their personal feelings about the question evaluated, and 5) that the words included in the item are unfamiliar. All these factors threaten the assumption of unidimensionality, and this violation implies a model misspecification that may lead to an incorrect ordering of thresholds (Andrich et al., 1997; Cheung & Mooi, 1994).

The first aim of this study is to test the assumptions of ordered response categories and ordered thresholds among response categories using response scales with three distinct middle categories. This will allow us to ascertain whether or not the central categories operate as expected from the integer scoring. The second aim is to determine whether a 3-point response scale is better than a dichotomous scale in terms of information and correlations with external criteria. To perform our study, we will use a set of items extracted from the 16PF-5 questionnaire (Cattell, Cattell & Cattell, 1993). On this questionnaire, all the items have a 3-point response scale in which the middle category is represented by a question mark (?). To broaden the scope of our study, we will also study the functioning of two additional middle response categories used in the former edition of the 16PF questionnaire (Cattell, Eber & Tatsuoka, 1970): «not sure» and «in between».

Our study differs from and extends previous research about middle response category in which IRT models were used. Andrich et al. (1997) and Rojas and Fernández (2000) used IRT models that assume that the response categories are ordered along the latent continuum, so that this assumption cannot be empirically tested. Just as Hernández et al. (2001) did, in this study we use

Bock's (1972) Nominal Model. This model does not impose any restrictions on the ordering of response categories, so that both assumptions of the integer scoring can be tested. Moreover, in the present study we investigate three middle response categories frequently used. We do not know any previous study in which «in between» and «?» have been investigated from this perspective.

Method

Sample

The sample of this study is composed of 816 under-graduate students who participated voluntarily. Data were gathered by means of a number of collective applications of the study scales. Those applications took place during the final part of a number of scheduled lectures. Regarding sex, 21.9% of the subjects were men. The average age was 22.19 (*s.d.* = 6.18).

Measures

We selected five items from factor H (social boldness) of the 16PF-5 Questionnaire (Cattell et al., 1993): items 9, 71, 107, 135 and 167. We decided to work with only 5 items as a compromise between: a) a sufficient number of indicators for tapping adequately the corresponding latent construct, and b) a small number of items and response categories compared to sample size, which justifies the use of the likelihood ratio chi-square statistic for testing model fit (Bock, 1997). On the 16PF-5 Questionnaire, responses are made along a 3-point response scale with middle category «?». On the basis of the integer scoring, a score of 0, 1 or 2 is assigned to each response category.

Two variables (anxiety and self-esteem) were used as external criteria in order to compare the correlations that the different response formats analyzed showed with them. These variables were used because they have showed significant correlations with the H factor in previous studies, and are used as external criteria in the questionnaire's manual (Cattell et al., 1993). Anxiety was measured by the trait anxiety subscale of the State-Trait Anxiety Inventory (Spielberger, Gorsuch & Lushene, 1970). This subscale is made up of 20 items that were responded to with a 4-point response scale. Self-esteem was measured with the 10-item scale developed by Rosenberg (1965). This scale was also responded to using a 4-point response scale. The alpha reliability estimates obtained for these scales were 0.89 and 0.77, respectively.

Design

The set of five items selected was presented to all the subjects with four different response scales. Three of them were polytomous response scales with 3 response categories. In one of them, the middle category was «?» (as in the 16PF-5 questionnaire), and the middle category for the other two polytomous response scales were «Not sure» and «In between» (response categories used in the previous 16PF edition). The fourth response scale was a dichotomous one. Thus, four 5-item scales with distinct response scales were presented to all the subjects: 1. $H_?$: an item scale with a polytomous response scale including «?» as the middle category; 2. H_{IB} : an item scale with a polytomous response scale including «In between» as the middle category; 3. H_{NS} : an item scale with a polytomous response scale including «Not sure» as the middle ca-

tegrity; and 4. H_{DIC} : an item scale with a dichotomous response scale.

Between every pair of the aforementioned scales, we included a different group of questions in order to decrease recall effects. Between the first scale and the second one, we included a number of questions referring to some demographic variables (sex, age, studies, reading habits, and languages used). Between the second and the third scales, we included the Rosenberg Self-Esteem Scale (Rosenberg, 1965). Lastly, between the third and the fourth scales, we included the Trait Anxiety Subscale of the State-Trait Anxiety Scale (Spielberger et al., 1970).

In order to control for the effects of the order of presentation of the four different scales, a latin square design was used to select four sequences of presentation among the 24 possible sequences. Approximately the same number of subjects responded to each of the four sequences. Within each scale, the order of items was randomly determined.

Analysis

We assessed the unidimensionality of the 5-item scales by means of a series of Confirmatory Factor Analyses (CFA) of the polychoric correlation matrix among the items with a polytomous response scale, and a CFA of the tetrachoric correlation matrix among the items with a dichotomous response scale. The Weighted Least Squares estimation method, as implemented in LISREL 8 (Jöreskog & Sörbom, 1993), was used.

To test the assumptions of ordered response categories and ordered thresholds, Bock's Nominal Model (BNM, Bock, 1972), as implemented in MULTILOG 6.0 (Thissen, 1991), was fitted to data for each scale of items with a polytomous response scale. In this model, each response category can be described by an option characteristic curve, and subject j 's probability of responding to the k^{th} category of the i^{th} item (P_{ijk}) can be expressed as follows:

$$P_{ijk}(\theta_j) = \exp(\theta_j a_{ik} + c_{ik}) / \sum_{h=1}^s \exp(\theta_j a_{ih} + c_{ih})$$

where item i has s response categories and $h = 1, 2, \dots, k, \dots, s$, θ_j is subjects j 's value on a continuous latent trait, and the a_{ik} and the c_{ik} are parameters associated with the k^{th} option of item i and are constrained to sum to 0 for each item i . a_{ik} can be interpreted as the discrimination between the k^{th} and the first response category (Mellenbergh, 1995). c_{ik} can be defined as the natural logarithm of the ratio between the probability of selecting the k^{th} category and the probability of selecting the first category when $\theta = 0$ (i.e., the value of the log odds at $\theta = 0$) (Mellenbergh, 1995). If the log odds is larger than zero, then, at $\theta = 0$, the person tends to prefer the k^{th} category over the first. If the a_{ik} values for item i are strongly ordered, then the corresponding response categories are ordered (Samejima, 1972; Bock, 1997). Thus, the examination of a_{ik} estimates allows us to test the assumption of ordered response categories. The thresholds (τ_{ik}) between two successive response categories must be estimated in order to test the assumption of ordered thresholds. In the BNM framework this can be done as follows (Bock, 1972, 1997):

$$\tau(\theta_{ik}) = (c_{ik} - c_{i,k-1}) / (a_{i,k-1} - a_{ik})$$

Results

Regarding unidimensionality, the goodness-of-fit indices obtained for the one-factor model were not satisfactory for the four scales of items (e.g., the Root Mean Square Error of Approximation (RMSEA) values were greater than 0.1). In every case, the modification index (MI) associated with the correlation between the errors of items 9 and 135 was very high ($MI_{09,135} = 67.24$ for the scale H_{NS} ; $MI_{09,135} = 49.79$ for the scale H_7 ; $MI_{09,135} = 53.93$ for the scale H_{IB} ; and $MI_{09,135} = 28.46$ for the dichotomous response scale (H_{DIC})). These results suggested a second dimension for items 9 and 135 that has to do with item wording: whereas items 9 and 135 were positively worded, the remaining items were negatively worded. To prevent this method factor that threatens unidimensionality, the item with the smallest factor loading in this pair (item 9) was removed from further analysis, and the resulting four 4-item scales were factor analyzed again. The goodness of fit indices obtained were acceptable (for the four 4-item scales: RMSEA = 0.08; Adjusted Goodness of Fit Index (AGFI) = 0.97; Non-Normed Fit Index (NNFI) = 0.94), therefore the assumption of a single dominant factor is tenable.

For each polytomous 4-item scale, the BNM was fitted to data. To assess model fit, MULTILOG yields the following statistic: negative twice the log likelihood estimated for the model ($-2\log\lambda$). This statistic is chi-square distributed on $(S-1) - 2n(r-1)$ degrees of freedom (Bock, 1997), where S is the number of response patterns, n is the number of items, and r the number of response categories. The $-2\log\lambda$ statistic is very sensitive to sample size, so that with large samples almost every model would be rejected (Hambleton, Swaminathan & Rogers, 1991). Thus, in the present study, we computed the ratio between $-2\log\lambda$ and the model's degrees of freedom as the basis for model fit assessment (Bock, 1997; Drasgow, Levine, Tsien, Williams & Mead, 1995; Gray-Little, Williams & Hancock, 1997). A ratio smaller than 3 is considered an indicator of an acceptable fit for the model (Bock, 1997; Drasgow et al., 1995). The BNM showed an acceptable fit for all the three polytomous scales ($-2\log\lambda/df$ equaled 1.84, 2.98 and 2.85, for the H_{NS} , H_{IB} , and the H_7 scales, respectively).

The a parameter estimates provided by the BNM are displayed in Table 1. The assumption of ordered response categories was supported in all the cases, regardless of which middle response category was involved. In all of these items, the a parameter estimates were ordered as expected from the integer scoring. However, only the H_{IB} scale showed ordered thresholds for all the items (see Table 1). This assumption was not satisfied in three items of scale H_7 (items 71, 107 and 135) and in one item of scale H_{NS} (item 107), although in this scale the two thresholds for item 167 showed almost the same value (-0.04 and -0.02).

Before comparing the dichotomous response scale with the polytomous ones, the 2-parameter logistic model was fitted to the data provided by the H_{DIC} scale by means of MULTILOG (see parameter estimates in Table 1). The model showed an acceptable fit to data ($-2\log\lambda/df = 1.314$). Because the only polytomous scale that satisfied the assumptions of ordered response categories and ordered thresholds was the scale with the central category «In between» (H_{IB}), the comparison was only carried out between this polytomous scale and the dichotomous scale (H_{DIC}). Firstly, we computed the information functions for both scales (see Figure 4). The results obtained showed that the polytomous scale (H_{IB}) provided more information than the dichotomous scale (H_{DIC}) for a wide range of θ values, with the largest absolute difference (1.52) obtained at $\theta = 0.5$.

Table 1
Parameter estimates yielded by Bock's Nominal Model and the 2-parameter logistic model

Response categories						Response categories					
SCALE H_{NS}		Yes	NS	No		SCALE H_{IB}	Yes	IB	No		
ITEM 71	a	-2.08	0.01	2.06		ITEM 71	a	-1.34	-0.04		1.38
	τ	-0.22	0.04		τ		-0.27	0.23			
ITEM 107	a	-1.39	0.11	1.29		ITEM 107	a	-2.21	-0.09		2.31
	τ	0.77	-0.10		τ		0.16	0.63			
ITEM 135	a	-0.84	0.14	0.69		ITEM 135	a	-0.98	0.12		0.86
	τ	-0.04	-0.02		τ		-0.46	0.71			
ITEM 167	a	-0.33	-0.11	0.44		ITEM 167	a	-1.31	0.21		1.10
	τ	-5.68	-1.40		τ		-0.75	0.71			

SCALE H_2		Yes	?	No	SCALE H_{DIC}		
ITEM 71	a	-1.27	-0.04	1.31	ITEM 71	a	1.48
	τ	0.10	-0.23			b	0
ITEM 107	a	-1.67	0.03	1.64	ITEM 107	a	1.46
	τ	0.41	0.32			b	0.37
ITEM 135	a	-0.88	0.05	0.83	ITEM 135	a	0.8
	τ	-0.02	-0.03			b	-0.12
ITEM 167	a	-1.19	0.09	1.10	ITEM 167	a	0.42
	τ	-0.42	0.27			b	-2.02

Note. NS: Not sure; IB: In between; DIC: dichotomous.

Secondly, we computed the correlations between the polytomous (H_{IB}) and the dichotomous (H_{DIC}) scales on the one hand, and both the anxiety and self-esteem measures on the other. The correlations obtained for the polytomous scale were -0.36 with anxiety and 0.33 with self-esteem. The correlations computed for the dichotomous scale were -0.34 and 0.33 , respectively. All the correlations were statistically significant ($p < 0.01$) and in the expected direction. The polytomous scale showed correlations slightly higher than the dichotomous one, but these differences were not statistically significant.

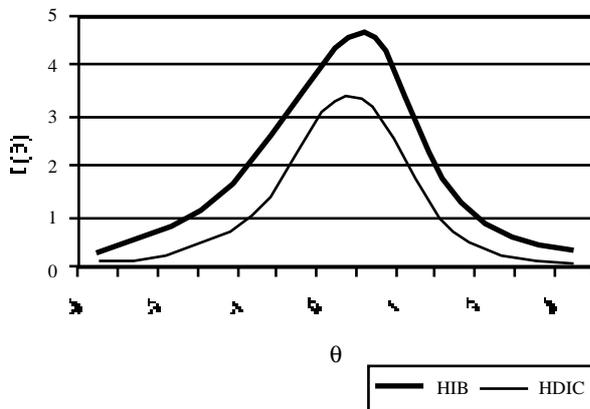


Figure 4. Information function for scales H_{IB} and H_{DIC}

Discussion

Many psychological tests and questionnaires are responded to by means of graded response scales, in which subjects' answers are scored according to the integer scoring. This scoring procedure assumes that response categories and the thresholds between them are ordered. However, there is no guarantee that response categories and thresholds will operate in the expected way (Andrich et al., 1997), so that their supposed ordering must be treated as hypotheses to be tested. The first aim of the present study was to test these assumptions by fitting Bock's (1972) Nominal Model (BNM) to data gathered from a sample of subjects who responded to a set of items extracted from the social boldness scale of Cattell et al.'s 16PF-5 Questionnaire.

The results obtained showed that only when the middle category of the response scale was «In between», was the assumption of ordered response categories and ordered thresholds supported. For the response scales that included «Not sure» or «?» as the middle response category, only the assumption of ordered response categories was supported. This means that in these two cases, the central category is not performing as expected in terms of its probability of being selected. That is, for subjects with intermediate levels in θ , the probability of responding by selecting the central category is smaller than the probability of responding by selecting other responses. These results are similar to those reported by Andrich and colleagues (1997) referring to the middle response category «Not sure», Rojas and Fernández (2000) referring to «Neither in agreement nor disagreement», and Hernández and

colleagues (2001) referring to «Indifferent». In all the items with disordered thresholds found by these researchers, the middle response category was involved.

The results obtained in the present study and those reported by previous investigations suggest that middle response categories that do not explicitly refer to an intermediate position in the latent construct (such as «Not sure», «Indifferent», and «?») generally are not interpreted by respondents as expected from the integer scoring, that is, as a middle position in the latent construct. And this might explain why these response categories are not performing as expected in terms of their probability of being selected. Middle response categories that do explicitly refer to an intermediate position in the latent construct (such as «In between») are probably interpreted by respondents as expected from the integer scoring, and that is why they function as expected. The results reported by Rojas and Fernández (2000) seem to support this idea. They found that for respondents who interpreted the middle response category «Neither in agreement nor disagreement» as an intermediate position in the latent trait, the assumption of ordered thresholds was supported and, consequently, there was an interval of θ values within which this response category was the category with the highest probability of being selected. For respondents who interpreted the aforementioned response category as an expression of doubt and indecision, the assumption of ordered thresholds was not supported.

Our results point out that «In between» should be preferred to «Not sure» and «?» as the middle response category in 3-point polytomous response scales. But does a 3-point scale including «In between» as the middle category yield any gains in comparison with a dichotomous response scale? This question was addressed by the second aim of our study. We compared both response scales in terms of their information functions and their correlations with external criteria. Regarding the information function, the results obtained showed that the polytomous format performed better than the dichotomous one along the latent construct continuum. Regarding correlations with external criteria, the results obtained showed that there were no statistically significant differences between the correlations that the dichotomous and the polytomous response scales had with the self-esteem and anxiety

measures. Taken together, these results suggest that the 3-point response scale with «In between» as the central category should be preferred to the dichotomous response scale.

The present study has a number of limitations. First, the study design might have contributed to the observed results. In those cases in which the item scale with the dichotomous response scale was presented to subjects before the other item scales, this could have led to subjects to consistently select a response category different from the central category when they responded to the item scales with a polytomous response scale. However, the fact that this possible limitation has not affected in the same way to the three polytomous response scales used in the present study, weakens this alternative explanation. Secondly, this design does not allow us to determine the causes of the disordering of thresholds. We have suggested that the interpretation of the middle response category may have one plausible explanation. This idea is related to multidimensionality of responses. As far as subjects' responses are caused by factors distinct from the latent construct measured (such as ambivalence, doubt, indecision, indifference; Dubois & Burns 1975), multidimensionality of responses is involved to a certain degree. Third, we assessed the unidimensionality of each scale by means of CFA. The results obtained pointed out that the hypothesis of a single dominant latent factor underlying the corresponding set of items could be maintained. However, the adequate fit of a model does not imply that it is the only 'true' model, and it does not remove the possibility of other secondary latent factors influencing subjects' responses. New studies with suitable designs should address this issue in the future. Finally, we did not analyze original item scales but rather abridged versions. Taking into account that the aim of our study is not to analyze a specific questionnaire, but instead to test the assumptions of the integer scoring when using the «Not sure», «?» and «In between» as the middle category, we think that this second limitation is not a problematic one.

Acknowledgements

The authors wish to thank two anonymous reviewers for their comments, and to Albert Maydeu and Bryce Reeves for their suggestions as to the analyses carried out.

Referencias

- Andrich, D. & Schoubroeck, L. (1989). The General Health Questionnaire: a psychometric analysis using latent trait theory. *Psychological Medicine*, 19, 469-485.
- Andrich, D. (1978a). A rating formulation for ordered response categories. *Psychometrika*, 43, 561-573.
- Andrich, D. (1978b). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement*, 2, 581-594.
- Andrich, D.; de Jong, J. & Sheridan, B.E. (1997). Diagnostic opportunities with the Rasch model for ordered response categories. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 58-68). Münster, Germany: Waxman Verlag.
- Bock, R.D. (1972). Estimating item parameters and latent ability when the responses are scored in two or more nominal categories. *Psychometrika*, 37, 29-51.
- Bock, R.D. (1997). The nominal categories model. In W. J. van der Linden & R. K. Hambleton (eds.), *Handbook of modern item response theory*. New York: Springer-Verlag.
- Cañadas, I. & Sánchez-Bruno, A. (1998). Categorías de respuesta en escalas tipo Likert (Response categories in Likert-type scales). *Psicothema*, 10, 623-631.
- Cattell R.B., Cattell, A.K.S. & Cattell H.E.P. (1993). *Handbook for the Sixteen Personality Factor Questionnaire*. Champaign, IL: Institute for Personality and Ability Testing.
- Cattell, R.B.; Eber, H.W. & Tatsuoka, M.M. (1970). *Handbook for the Sixteen Personality Factor Questionnaire (16 PF)*. Champaign, IL: Institute for Personality and Ability Testing.
- Cheung, K.C. & Mooi, L.C. (1994). A comparison between the rating scale model and dual scaling for Likert scales. *Applied Psychological Measurement*, 18, 1-13.
- Cronbach, L.J. (1946). Response sets and test validity. *Educational and Psychological Measurement*, 6, 475-494.
- Cruickshank, P.J. (1984). A stress and arousal mood scale for low vocabulary subjects: a reworking of Mackay et al. (1978). *British Journal of Psychology*, 75, 89-94.

- Drasgow, F.; Levine, M.V.; Tsien, S.; Williams, B. & Mead, A.D. (1995). Fitting polytomous item response theory models to multiple choice tests. *Applied Psychological Measurement*, 19, 143-165.
- 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.
- Edwards, A.L. (1946). A critique of «neutral» items in attitude scales constructed by the method of equal appearing intervals. *Psychological Review*, 53, 159-169.
- Goldberg, G. (1971). *Response format in attitude scales*. Unpublished manuscript, Northwestern University.
- Gray-Little, B.; Williams, V.S.L. & Hancock, T.D. (1997). An items response theory analysis of the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin*, 23, 443-451.
- Hambleton, R.K.; Swaminathan, H. & Rogers, H.J. (1991). *Fundamentals of Item Response Theory*. Newbury Park: SAGE.
- Hernández, A.; Espejo, B.; González-Romá, V. & Gómez-Benito (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.
- Joreskog, K.G. & Sörbom, D. (1993). *LISREL 8 user's reference guide*. Chicago, IL: Scientific Software International, Inc.
- Kaplan, K. (1972). On the ambivalence-indifference problem in attitude theory: A suggested modification of the semantic differential technique. *Psychological Bulletin*, 77, 361-372.
- Mellenbergh, G. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement*, 19, 91-100.
- Rojas, A.J. & Fernández, J.S. (2000). Análisis de las alternativas de respuestas intermedias mediante el modelo de escalas de clasificación (Analysis of response categories by means of the rating scale model). *Metodología de Encuestas*, 2, 171-183.
- Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.
- Samejima, F. (1972). A general model for free-response data. *Psychometrika Monograph*, No. 18.
- Spielberger, C.D.; Gorsuch, R.L. & Lushene, R.E. (1970). *STAI. Manual for the State-Trait Anxiety Inventory*. Palo Alto, CA: Consulting Psychologists Press.
- Thissen, D. (1991). *Multilog User's Guide*. Chicago: Scientific Software.
- Worthy, M. (1969). Note on scoring midpoint responses in extreme response style scores. *Psychological Reports*, 24, 189-190.
- Wright, B.D. & Masters, G.N. (1982). *Rating scale analysis*. Chicago: Mesa Press.