

The functioning of central categories *Middle Level* and *Sometimes* in graded response scales: Does the label matter?

Ana Hernández, Begoña Espejo and Vicente González-Romá
Universidad de Valencia

The present study evaluates the extent to which central categories explicitly labeled as being in the middle of the other response categories, specifically *Middle Level* and *Sometimes*, function as expected according to the integer scoring system. The assumptions are tested by means of Bock's Nominal Model in two 5-response scales. Results show that the assumption of the ordering of the response categories is met for all the items. The ordering of thresholds is satisfied for all but one item with the central category *Middle Level*. Results are compared with those obtained when middle categories are not explicitly labeled as being in the middle of the other response categories, as in the case of *Not Sure*, *Undecided* or?

Funcionamiento de las categorías centrales Término medio y A veces en escalas de respuesta graduada: ¿es importante el anclaje verbal? En este estudio se evalúa en qué medida las categorías centrales *Término medio* y *A veces*, explícitamente etiquetadas para reflejar una posición intermedia respecto del resto de categorías de la escala, funcionan como se asume desde el sistema de puntuación habitual en este tipo de escalas. El cumplimiento de los supuestos del sistema de puntuación se evalúa mediante el Modelo nominal de Bock en dos escalas con 5 alternativas de respuesta. Los resultados muestran que, si bien el orden de las categorías de respuesta se mantiene para todos los ítems, los umbrales entre las categorías de respuesta se desordenan para uno de los ítems con categoría central *Término medio*. Los resultados son comparados con los obtenidos cuando las categorías centrales no están explícitamente etiquetadas para reflejar una posición central, como es el caso de las categorías *No estoy seguro*, *Indeciso* o?

Items with polytomous graded response scales are frequently used for measuring psychological constructs. Subjects respond to these kinds of items by selecting the response category that best represents their position on the latent construct. Successive integers are assigned to the successive response categories based on the assumption that the integers approximately reflect the respondent's standing on the underlying latent construct. This scoring procedure is called *integer scoring* (e.g., Andrich and Schoubroeck, 1989).

Frequently, polytomous graded scales present a central response category that is used to represent an intermediate stand on the latent construct. The results obtained in a number of studies (Andrich, de Jong and Sheridan, 1997; Asensio and Rojas, 2002; González-Romá and Espejo, 2003; Hernández, Espejo, González-Romá and Gómez, 2001) suggest that central categories with labels that are not explicitly assigned an intermediate position in the latent construct, such as *Not Sure*, *Indifferent*, *Neutral*, *Undecided* or?, are not generally used by respondents in the way assumed by the integer scoring. Some of the studies also suggest that central

categories with labels that explicitly refer to an intermediate position on the latent construct that lies between the other response categories, such as *In Between*, do function as assumed by the integer scoring (González-Romá and Espejo, 2003).

This study investigates whether frequency and quantitative response scales (Cañadas and Sánchez-Bruno, 1998; Durán, Ocaña, Cañadas and Pérez-Santamaría, 2000) with the central categories *Sometimes* and *Middle Level*, which explicitly refer to an intermediate position between the other response categories, fulfill the integer scoring assumptions. These kinds of central categories are expected to more adequately represent an intermediate position on the latent continuum than categories like *Undecided* or *Not Sure* (Morales, 2000).

The integer scoring assumptions

The integer scoring system relies on two main assumptions (see González-Romá and Espejo, 2003). First, it is assumed that the m_j response categories are *ordered* along the continuous latent variable (θ) in the manner indicated by the integers assigned to each response category. So, for a three-response category item (e.g. «0. Disagree», «1. Middle level», «2. Agree»), people who show the lowest levels on θ will have the greatest probability of choosing the lowest category. People with intermediate levels on θ will be the ones with the greatest probability of choosing the middle response category. Finally, people with the highest levels

on θ will have the greatest probability of choosing the highest category. Second, it is also assumed that there are m_i-1 ordered thresholds (τ) that designate boundary positions between adjacent response categories, in such a way that within a given interval of θ values every response category has the highest probability of being selected. For example, for the three response category item, the most probable response will be 0 if $\theta < \tau_1$; the most probable response will be 1 if $\tau_1 < \theta < \tau_2$, and the most probable response will be 2 if $\theta > \tau_2$.

The probability of each subject's response to an item is determined by the interval in which the subject's value in θ falls. Therefore, when the two aforementioned assumptions are met, the score obtained by a subject by means of integer scoring represents the number of ordered thresholds the subject has passed (Andrich, and Schoubroeck, 1989). Figure 1 (Panel A) shows the Category Response Functions (CRFs) for a 3-category item when both assumptions are met.

Clearly, when the first assumption is not met, the CRFs will not appear in the expected order. Subjects with a specific position in θ will not have the greatest probability of choosing the category that is expected according to integer scoring (see the example depicted in Figure 1, Panel B). When the response categories are ordered as assumed by the integer scoring system, but the thresholds between them are not, the second assumption is not met. In this case, even though people who have the greatest probability of choosing a category are those that one would expect, based on their position on the latent continuum, they will show an even greater probability of choosing one of the other categories. For the example depicted in Figure 1 (Panel C), it can be seen that the middle category is more likely to be chosen by people in the middle of the continuum, but these people have a greater probability of choosing either of the other two categories. Consequently, when thresholds are not ordered, at least one of the categories involved is not performing as expected in terms of its probability of being selected. There is no interval on the θ continuum for which that category has the highest probability of being selected. In this case, the score obtained by a subject does

not represent the number of ordered thresholds passed by the subject.

Testing the integer scoring assumptions with different middle categories: empirical evidence

Some studies have tested the assumption of ordered thresholds by using the so-called Rasch models (Andrich et al, 1997; Asensio and Rojas, 2002; Rojas and Fernández, 2000). Others have tested both the ordering of categories and the ordering of thresholds by using Bock's Nominal Model (NM, Bock, 1972) (González-Romá and Espejo, 2003; Hernández et al, 2001).

One of the first studies to test the hypothesis of ordered thresholds was conducted by Andrich et al (1997). They analyzed a 5-point-Likert-response questionnaire with a middle category labeled as *Not Sure*. Results showed that many items did not show ordered thresholds. In all cases, the disordered-threshold problem involved the middle category, which did not work as a category in the middle of the others. Andrich and colleagues (1997) concluded that their 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 not be treated as an attitude more or less somewhere between a negative and a positive attitude» (Andrich, et al, 1997, p. 66).

More recently, Asensio and Rojas (2002) compared the functioning of a 5-point Likert response scale with different middle response category labels and formats. Specifically they compared five forms of the scale. For three of the forms all response options were numbered (from 1 to 5) and a different middle category label was used in each form; *Indifferent*, *Undecided* and *Neither in agreement nor disagreement*. For another form, the middle category *Neither in agreement nor disagreement* was used again, but no number accompanied any of the response options in this case. Finally, a response scale numbered from 1 to 5 with no labels accompanying the response options was also evaluated. After fitting a Rasch model for Rating Scales (Andrich, 1978; Masters, 1980), results showed that only

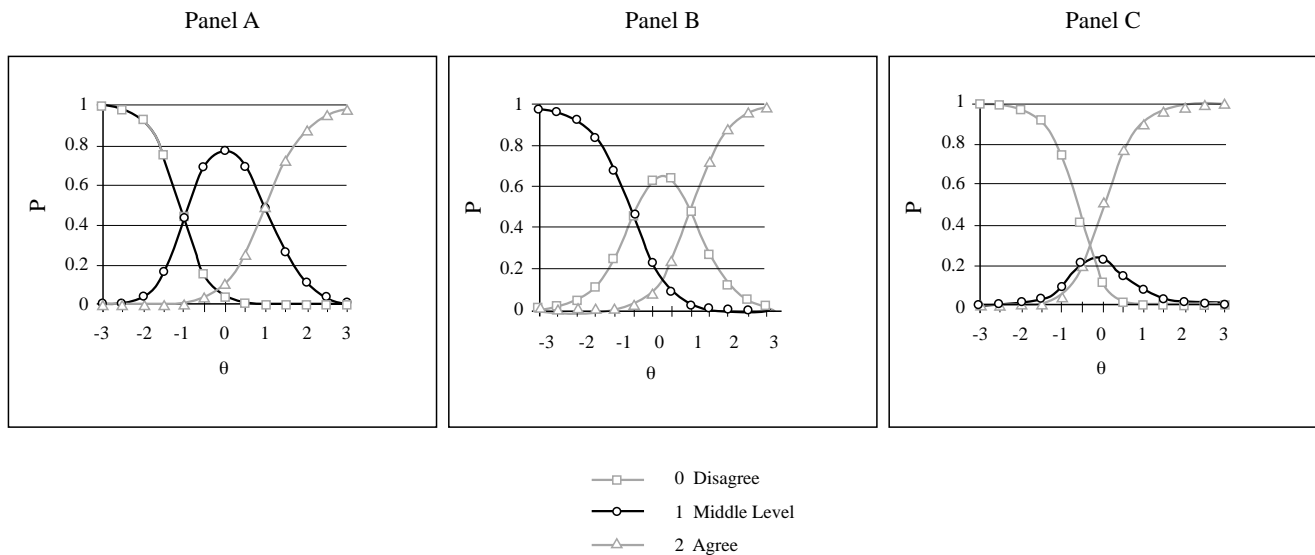


Figure 1. Response functions for an item with three response categories. Panel A: ordered categories and ordered thresholds. Panel B: Disordered categories. Panel C: Disordered thresholds

the scale with the numeric middle category 3 that had not a specific label showed ordered thresholds. For the remaining scales, the assumption of ordered thresholds was not supported. When the scales with middle categories *Undecided*, *Indifferent* or *Neither in agreement nor disagreement* were used, the middle category did not work as a category in the middle of the others.

Regarding the studies that tested both the hypothesis of ordered response categories and the hypothesis of ordered thresholds, Hernández et al (2001) and González-Romá and Espejo (2002) evaluated a number of 5-point and 3-point Likert response scales with central categories labeled as *Indifferent*, *Not Sure*, *?*, and *In Between*. Their results showed that the assumptions of the integer scoring were met for all the items only when the middle response category was *In Between*. For the scales with central categories *Indifferent*, *Not Sure* or *?*, the assumption of ordered categories was met for all the cases, but the assumption of ordered thresholds was not met for most of the items. Once again, the central category was the one involved. There was no interval on the latent continuum for which the central category had the highest probability of being selected.

Thus, it is clear that response scales that include middle categories frequently do not function as expected using the integer scoring system, since the assumption of ordered thresholds is not met. Some studies suggest that threshold disordering might depend on the meaning subjects assign to the central category. Specifically interpreting the central categories in a different way from the expected «having an intermediate position on the continuum» (e.g. to show ambivalence or indifference, as a response style, as a sign of doubt, as a negative to reveal personal feelings, etc) is a potential source of multidimensionality that can cause the disordering of the thresholds (Andrich, et al, 1997; Cheung and Mooi, 1994). Along these lines, Rojas and Fernández (2000) evaluated the ordering of thresholds on a 5 point scale with a middle category labeled *Neither in agreement nor disagreement*. Their results showed that the assumption was only supported when subjects assigned to the middle response category the meaning of «an intermediate position in the latent trait».

The present study

The results previously presented suggest that the label of the middle response category could play an important role in its functioning. The aim of this study is to ascertain whether frequency and quantitative response scales with central categories *Sometimes* and *Middle level*, explicitly labeled as being in the middle of the other response categories, meet the integer scoring assumptions.

Method

Sample

The sample of this study was made up of 932 employees of a Spanish regional public health service who participated in a research project about quality of life at work. A two-stage randomized sampling procedure was carried out. In the first stage, 250 work units were randomly selected. In the second stage, four subjects were selected from each unit: the supervisor and three members who were randomly selected. The average age of the total sample was 41.18 years ($SD= 9.18$), and the average tenure

in the regional public health service was 14.58 years ($SD= 7.9$). Regarding gender, 37.9% of the sample were men.

Measures

Two scales were used. The first one was made up of 4 of the 6 items included in the organizational demands scale constructed by Schaufeli, González-Romá, Peiró, Geurts and Tomás (in press). Respondents answered using a response scale ranging from «1. Not at all» to «5. Very much», with a central response category labeled «3. Middle level» («Término medio» in the original Spanish Questionnaire). The reason only 4 out of the 6 items were selected was that the inclusion of all the items resulted in a number of possible response patterns ($5^6= 15625$) that exceeded the number of response patterns that could be observed in our sample of 932 subjects. The data would have been very sparse, and the expected frequencies would have been extremely small. Therefore, the application of the classical goodness of fit indices based on the chi-square distribution would have been inappropriate. Consequently, only the 4 items with the highest factor loadings were chosen. The alpha coefficient for the 4-item scale was 0.80. The second scale was made up of 3 items that were selected on their face validity from the emotional exhaustion scale of the Maslach Burnout Inventory (Maslach and Jackson, 1986). This short scale has been shown to have good psychometric properties (see Peiró, González-Romá, Tordera and Mañas, 2001). In this case, the response scale ranged from «1. Never» to «5. Many times», with a central category labeled «3. Sometimes» («A veces» in the original Spanish Questionnaire). The alpha coefficient for this 3-item scale was 0.79.

Analysis

To assess the dimensionality of both scales we carried out two separate principal component analyses by means of PRELIS 2 (Jöreskog and Sörbom, 1993). The polychoric correlation matrices among the involved items were analyzed and the ordered eigenvalues and percentages of variance explained by the components were evaluated (e.g., Barr and Raju, 2003; González-Romá, Hernández and Gómez-Benito, in press). Then, the NM (Bock, 1972) was fitted to both scales by means of Multilog 6.0 (Thissen, 1991). In this model, subject j 's probability of responding to the k_{th} category of the i_{th} item (P_{ijk}) can be expressed as follows:

$$P_{ijk}(\theta_j) = \exp(\theta_j a_{ik} + c_{ik}) / \sum_{k=1}^m \exp(\theta_j a_{ik} + c_{ik})$$

where item i has m response categories and $k= 1, 2, \dots, k, \dots, m$. 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 . If the a_{ik} values for item i are strongly ordered, then the corresponding response categories are ordered (Samejima, 1972; Bock, 1997). Thus, 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 NM this can be done as follows (Bock, 1972, 1997):

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

To assess the model's goodness-of-fit, MULTILOG yields the following statistic: negative twice the log-likelihood estimated for the model (-2logλ). 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 is the number of response categories. The -2logλ statistic is very sensitive to sample size. With large samples, almost every model would be rejected (Hambleton, Swaminathan and Rogers, 1991). Thus, in the present study, we computed the ratio between -2logλ and the model's degrees of freedom as the basis for model fit assessment (Drasgow, Levine, Tsien, Williams and Mead, 1995; Bock, 1997; Gray-Little, Williams and Hancock, 1997). A ratio smaller than 3 is considered to be an indicator of a satisfactory fit for the model (Bock, 1997; Drasgow et al, 1995).

Results

Dimensionality

For each scale only one component with an eigenvalue greater than 1 was obtained. The first component of the «organizational demands» scale explained 67.84% of the variance and the first component of the «emotional exhaustion» scale explained 74.67% of the variance. So, we concluded that the two scales met the assumption of unidimensionality required to apply the NM.

Fitting the Nominal Model

For each scale, the NM was fitted to data. In both cases, the model showed an acceptable fit according to the criteria proposed by Bock (1997) and Drasgow et al, (1995): the ratio between -2logλ and the model's degrees of freedom was smaller than 3 (see table 1).

The a parameter estimates provided by the NM for the items on both scales are displayed in table 2. For the items that make up the «organizational demands» scale with a central category labeled *Middle Level*, the assumption of ordered response categories was supported in all cases. The a parameter estimates were ordered as expected from the integer scoring for the four analyzed items. The assumption of ordered thresholds (τ) was satisfied in all the items but one (item 4). In figure 2 we show the CRFs for one of the items with ordered response categories and ordered thresholds, specifically for item 2. It is observed that all the CRFs are ordered along the θ continuum, and all show an interval in θ for which they have the highest probability of being selected compared to any of the other categories. In figure 3, we show the CRFs for the item in which the assumption of ordered thresholds was not supported

(item 4). In this figure, we can see that, although the five CRFs are ordered along the θ continuum, for some of the categories there is no interval on θ for which they have the highest probability of being selected compared to the other categories. Specifically, there is no range of θ values within which the middle response category *Middle Level* and the fourth category *Quite a Lot* show the highest probability of being selected.

Regarding the items that compose the «emotional exhaustion» scale, with a middle response category labeled *Sometimes*, both the assumption of ordered response categories and the assumption of ordered thresholds were supported for all the items, since both the a and τ parameter estimates were ordered along the latent continuum. The representation of the CRFs in this case was similar to that depicted in figure 2.

Table 2
a parameter estimates yielded by the Nominal Model, and threshold estimates (τ) for both scales

Organizat. demands	Response categories				
	Not at all	Just a little	Middle level	Quite a lot	A lot
Item 1	a -2.08	-0.90	0.15	0.96	1.88
τ	-0.95	-0.69	-0.30	0.35	
Item 2	a -4.50	-1.70	0.86	2.06	3.28
τ	-1.19	-0.49	0.43	0.98	
Item 3	a 2.78	-0.73	0.16	0.40	0.93
τ	-0.83	0.08	0.40	0.93	
Item 4	a -1.90	-0.39	0.31	0.47	1.51
τ	-0.51	0.51	0.56	0.36	

Emotional exhaustion	Never	Occasionally	Sometimes	Much of the time	Most of the time
Item 1	a -2.17	-1.13	-0.09	1.17	2.22
τ	-2.65	-1.88	-0.03	0.83	
Item 2	a -5.93	-3.62	-0.10	3.43	6.22
τ	-1.48	-0.53	0.36	0.98	
Item 3	a -2.75	1.43	0.14	1.75	2.29
τ	-0.78	.29	1.39	2.63	

Table 1
Fit indices for Bock's Nominal Model in each of the analyzed scales

	Scales	
	Organizational demands	Emotional exhaustion
-2logλ	796.1	93.5
gl	273	101.0
-2logl/gl	2.92	0.93

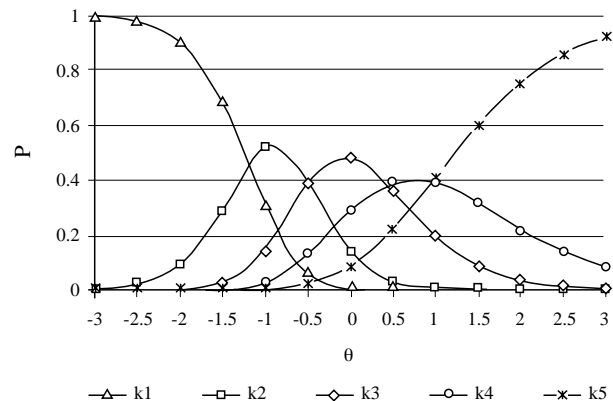


Figure 2. Category response functions for item 2 (Organizational demands), with ordered response categories and ordered thresholds

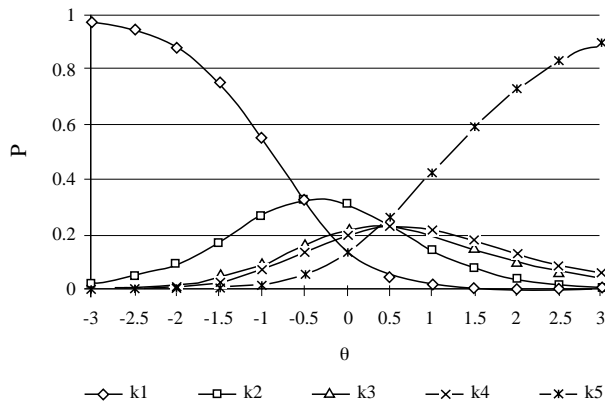


Figure 3. Category response functions for item 4 (*Organizational demands*), with ordered response categories and disordered thresholds

Discussion

Items with polytomous graded response scales are usually scored following the integer scoring system. This scoring procedure assumes that both the response categories and the thresholds between them are ordered. When these assumptions are met, a subject's score reflects the number of ordered thresholds he or she passes. However, these assumptions should not be taken for granted (Andrich et al. 1997). A number of studies (Andrich et al. 1997; Asensio and Rojas, 2002; González-Romá and Espejo, 2003; Hernández et al. 2001) have shown that central categories with labels not explicitly referring to an intermediate position in the latent construct, such as *Not Sure*, *Indifferent*, *Undecided* and *?*, are not generally used by respondents as assumed in the integer scoring, since the assumption of ordered thresholds is not met. Specifically, even though people with intermediate levels on the latent continuum are those who have the greatest probability of choosing the middle response category, they show an even greater probability of choosing one of the adjacent categories.

The present study evaluates the extent to which central categories explicitly labeled as being in the middle of the other response categories, specifically *Middle Level* and *Sometimes*, function as expected according to the integer scoring system. The assumptions are tested by means of Bock's NM (1972) in two 5-response scales.

Results show that the first assumption (the ordering of the response categories) is met for all the items in the two scales analyzed (see a estimates in table 2). So, as expected when applying the integer scoring, the probability of choosing a higher category increases as the latent continuum values increase. The second assumption (the ordering of the thresholds) is satisfied for the 3 items that make up the «Emotional Exhaustion» Scale, but it is not satisfied for one of the 4 items (item 4) on the «Organizational Demands» Scale (see τ estimates in table 2). For this item, the disordering of thresholds involves not only the middle category *Middle Level*, but also the fourth category *Quite a Lot*. For both categories, there is no interval in θ for which they have the highest probability of being selected (see figure 3).

Comparison of these results with those obtained in prior studies where middle categories were not explicitly referred to an intermediate position on the latent construct suggests that, in general terms, using middle categories explicitly labeled to express an intermediate position on the latent continuum, in

comparison with the other categories, can contribute to reducing the number of items with disordered thresholds. According to our results, this is especially true for the frequency scale with the central category *Sometimes*. Results obtained by González-Romá and Espejo (2003) with the central category *In Between* also support the better functioning of central categories explicitly referring to an intermediate position.

The use of midpoint labels that are close to the theoretical meaning expected in the integer scoring procedure can contribute to using and interpreting the response scales according to that theoretical meaning. However, the use of these kinds of labels does not guarantee that they will function as expected according to the integer scoring system. To take an obvious example, one of the items on the «Organizational Demands» scale showed disordered thresholds, in spite of the fact that the middle category was explicitly labeled as being in the middle of the other categories. So, although these kinds of labels seem to be more appropriate than more ambiguous labels, such as *Not Sure* or *?*, the use of explicitly middle labeled categories does not guarantee their appropriate use and interpretation. Middle categories can be selected for reasons other than the position of the individuals in the latent continuum. People can choose the middle category as a response tendency, as an expression of doubt or indecision, due to indifference or ambivalence, because they do not understand the statement, because they do not want to reveal their personal feelings, etc. (e.g., Dubois and Burns, 1975; Raaijmakers, van Hoof, Hart, Vergot and Vollebergh, 2000; Moustaki and O'Muircheartaigh, 2002; Presser and Schuman, 1980; Schuman and Presser, 1981). And all these possible interpretations can operate, even if the middle categories are labeled with the aim of maximizing their appropriate use by the individuals.

So, apart from using middle categories that explicitly refer to an intermediate position on the latent construct, some other strategies can be used in order to foster the appropriate use of the middle response categories. One possible strategy is to provide a second non-directional category labeled *Don't Know*, which is likely to attract individuals' responses that do not reflect their positions on the latent continuum. In this sense, Harter (1997) showed that adding a *Don't Know* alternative to a 5 point-Likert response scale improved the operating characteristics of the central category. A different strategy involves giving specific instructions about the use of the middle category, in such a way that people choose the middle category only if they feel they are located in the middle of the continuum compared to the other response categories. Results obtained by Rojas and Fernández (2000) support the usefulness of this strategy. Specifically, the assumption of ordered thresholds was supported when subjects were instructed to select the central category to reflect an intermediate position in the latent trait, and it was not supported when subjects were instructed to select the central category as an expression of doubt and indecision. Future studies should investigate the efficacy and relative importance of each of these alternative strategies. Finally, it is also possible to disregard central categories and use an even number of response options. However, different authors recommend including a central category, so people are not forced to choose an alternative that may not describe them (Neumann, 1979; Sudman and Bradburn, 1989).

Apart from the labeling of the middle categories, some other factors such as the number of response options could also play an

important role in the functioning and interpretation of the different response categories (Andrich and Master, 1988; Komorita and Graham, 1965; Weng, 2004). The impact of the number of response options on reliability, validity and model fit has been evaluated in a number of studies (e.g., Hernández, Muñiz and García-Cueto, 2000; Tomás and Oliver, 1998; Sancerni, Meliá and González-Romá, 1990). The possible impact of the interaction between the number of response categories and the label of the middle response option should be evaluated in future research.

The present study has a number of limitations we would like to point out. First, since the scales evaluated are measures of different constructs, the items that make up each scale are not the same. Ideally, the same items should have been measured with

different midpoints (as in the study carried out by González-Romá and Espejo, 2003). However, the results obtained in the present study still represent valuable empirical evidence accumulated for items with middle categories that explicitly refer to an intermediate position. Second, we have no information about the meaning subjects assigned to the middle categories *Sometimes* and *Middle Level*. So we cannot ensure that the correct interpretation of these central categories (according to the meaning expected under the integer scoring) is directly responsible for the observed ordered thresholds. The extent to which the meanings attributed to the different middle categories vary with the different labels and affect the ordering of thresholds differently should be addressed in future studies.

References

- Andrich, D., & Masters, G.N. (1988). Rating scale analysis. In J.P. Keeves (Ed) *Educational research, methodology, and measurement: An international handbook* (pp. 297-303). Oxford, UK: Pergamon
- Andrich, D., & Schoubroeck, L. (1989). The General Health Questionnaire: a psychometric analysis using latent trait theory. *Psychological Medicine*, 19, 469-485.
- 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*. Münster, Germany: Waxman Verlag, pp. 58-68.
- Asensio, M. & Rojas, A. (2002). Análisis de la categoría central de los ítems en función de su denominación mediante el modelo de escalas de clasificación. *Metodología de las Ciencias del Comportamiento, volumen especial*, 55-59.
- Barr, M.A. & Raju, N.S (2003). IRT-based assessments of rater effects in multiple-source feedback instruments. *Organizational Research Methods*, 6, 15-43.
- 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. En W. J. van der Linden, and R. K. Hambleton (Eds.): *Handbook of modern item response theory*. New York Inc.: 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.
- 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), 1-13.
- Dragow, 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, (2), 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.
- Durán, A., Ocaña, A. C., Cañadas, I., & Pérez-Santamaría, F. J. (2000). Construcción de cuestionarios para encuestas: el problema de la familiaridad de las opciones de respuesta (Survey construction: the problem of the familiarity of the response options). *Metodología de Encuestas*, 2, 27-60.
- Gonzalez-Romá, V. & Espejo, B. (2003). Testing the middle response categories »Not sure«, «In between» and «?» in polytomous items. *Psicothema*, 15, 278-284.
- Gonzalez-Romá, V., Hernández, A. & Gómez-Benito (in press). Power and Type I error of the mean and covariance structure analysis model for detecting differential item functioning in graded response items. *Multivariate Behavioral Research*.
- Gray-Little, B., Williams, V. S. L., & Hancock, T. D. (1997). An item response theory analysis of the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin*, 23, (5), 443-451.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park: SAGE.
- Harter, J. K. (1997) *The psychometric utility of the midpoint on a Likert scale*. Dissertation Abstracts International: 58, 1198.
- Hernández, A.; Muñiz, J. & García_Cueto, E. (2000) Comportamiento del modelo de respuesta graduada en función del número de categorías de la escala. *Psicothema*, 12, Sup. 2, 288-291.
- 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 relevant the «indifferent» alternative?). *Metodología de Encuestas*, 3, 135-150.
- Jöreskog, K. G., & Sörbom, D. (1993). *PRELIS 2 User's reference guide*. Chicago, IL: Scientific Software International, Inc.
- Komorita, S.S. & Graham, W.K. (1965) Number of scale points and the reliability of scales. *Educational and Psychological Measurement*, 15, 987-995.
- Maslach, C. & Jackson, S.E. (1986). *Maslach Burnout Inventory: Second edition*. Palo Alto, CA: Consulting Psychologists Press
- Morales, P. (2000). *Medición de actitudes en psicología y educación : construcción de escalas y problemas metodológicos*. 2nd Edition. Madrid: UPCCO publicaciones.
- Moustaki, I., & O'Muircheartaigh, C. (2002). Locating 'Don't Know', 'No Answer' and middle alternatives on an attitude scale: A latent variable approach. In Marcoulides, G. A. & I. Moustaki (Eds.), *Latent variables and latent structure models* (pp. 15-40). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Neumann, L. (1979). *Effects of categorization on relationships in distributions and applications to rating scales*. Dissertation Abstracts International, 40, 2262-B
- Peiró, J.M., González-Romá, V., Tordera, N & Mañas, M.A. (2001). Does role stress predict burnout over time among health care professionals? *Psychology & Health*, 16, 511-525.
- Presser, S. & Schuman, H. (1980). The measurement of a middle response option with opinion questions. *Public Opinion Quarterly*, 44, 79-85.
- Raaijmakers, Q. A. W., van Hoof, A., Hart Verbogt, T. F. M. A., & Vollebergh, W. A. M. (2000). Adolescents' midpoint responses in Likert-type scale items: neutral or missing values? *International Journal of Public Opinion Research*, 12, 208-216.
- Rojas, A., & Fernández, J. S. (2000). Análisis de las alternativas de respuestas intermedias mediante el modelo de escalas de clasificación (Analysis of the middle response options with the Rating Scale Model). *Metodología de Encuestas*, 2, 171-183.
- Samejima, F (1972). A general model for free-response data. *Psychometrika Monograph*, 18.
- Sancerni, M. D.; Meliá, J.L. & González-Romá, V. (1990) Formato de respuesta, fiabilidad y validez en la medición del conflicto de rol. *Psicológica*, 11, 167-175.
- Schaufeli, W.B., González-Romá, V., Peiró, J.M., Geurts, S. & Tomás, I. (in press). Withdrawal and burnout in health care: On the mediating role of inequity. In Korunka, C. & Hoffmann. P. (in print). Change and

- Quality in Human Service Work (Organizational Psychology and Health Care, Volume 4, Series editors: J. M. Peiro & W. Schaufeli). München: Hampp Publishers.
- Schuman, H. & Presser, S. (1981). Questions and answers in attitude surveys. New York; Academic Press.
- Sudman, S., & Bradburn, N. M. (1989). Measuring attitudes: recording responses. In S. Sudman & N. M. Bradburn (Eds.). *Asking questions: A practical guide to questionnaire design*. San Francisco: Jossey-Bass Publishers.
- Thissen, D. (1991). *MULTILOG user's guide. Multiple, categorical item analysis and test scoring using item response theory (Version 6.0). Computer program*. Chicago IL: Scientific Software International, Inc.
- Tomás, J.M. & Oliver, A. (1998) Efectos de formato de respuesta y método de estimación en el análisis factorial confirmatorio. *Psicothema*, 10, 197-208.
- Weng, L. (2004) Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability- *Educational and Psychological Measurement*, 64, 956-972.