Laboratorio de Evaluación
Psicológica y Educativa

Evaluar, 12 (2012), 43 – 59
ISSN 1667-4545

The Effect of the Number of Answer Choices on the Psychometric Properties of Stress Measurement in an Instrument Applied to Children

González-Betanzos Fabiola[1]*, Leenen Iwin**, Lira-Mandujano Jennifer* &
Vega-Valero Zaira***

*Facultad de Psicología, Universidad Michoacana de San Nicolás de Hidalgo, Morelia, Michoacán, México.

**Facultad de Medicina, Universidad Nacional Autónoma de México, México, DF.

***Facultad de Psicología, Universidad Nacional Autónoma de México, Tlalnepantla, México.

Abstract:
The main objective of this study was to use Item Response Theory (IRT) models to measure the effect exerted by the number of response options on the psychometric properties of a test measuring stress in children. In this study, we applied the 30-item Child Stress Perception Inventory (CSPI) scale to 583 children; the items have different response alternatives (3, 5, or 7). We studied whether the scales measure the same trait and whether the alternatives that the same items possess are equivalent. As evidence of validity, we present measurements that examine the internal structure of the instrument and its relationship with other variables. The result indicates that the three forms measure the same trait, but that there is no equivalency among the categories. The scale adjustment of 7 response alternatives is best; however, validity in relation to other variables is optimal for 5 response alternatives, which in addition, performs best in terms of reliability and information.

Key Words: Response format, IRT, Psychometric Properties, Child Stress Perception.

El efecto del número de opciones de respuesta sobre las propiedades psicométricas de la medida de estrés con un instrumento aplicado a niños

Resumen.
El presente trabajo tiene como principal objetivo analizar mediante modelos de la Teoría de Respuesta al Ítem (TRI) el efecto que tiene el número de alternativas de respuesta sobre las propiedades psicométricas de un test que mide estrés infantil. En el presente estudio se aplicó la escala de "Percepción de Estrés para Niños" (IPEI) de 30 ítems a 583 niños, los ítems tenían diferentes alternativas de respuesta (3, 5 o 7). Se estudio si las escalas miden el mismo rasgo y si las alternativas que tienen las mismas etiquetas son equivalentes. Como evidencias de validez se presentan medidas que examinan la estructura interna del instrumento y su relación con otras variables. Los resultados indican que las tres escalas miden el mismo rasgo pero no existe equivalencia entre las categorías. El ajuste de escala de 7 alternativas de respuesta es mejor, sin embargo, la validez en relación con otras variables es óptima para 5 alternativas de respuesta, que además muestra el mejor comportamiento en términos de fiabilidad e información.

Palabras clave: Formato de respuesta, TRI, propiedades psicométricas, percepción de estrés en niños.

Research on the optimal number of alternatives in scales that measure attitudes or personality has a long history that dates back to the 1920s. The majority of these works

[1] La correspondencia relacionada con este artículo debe enviarse a: Fabiola González Betanzos, Facultad de Psicología, Universidad Michoacana de San Nicolás de Hidalgo Francisco Villa 450 58110- Morelia, México. E-mail: betanzos@umich.mx. Teléfono: 52 4433129913. Fax : 52 443312991.

conclude that the number of alternatives conditions, to a greater or lesser degree, the response of those being examined to an item. Thus, the choice of possible response alternatives and their interaction with the measurement of the construct has become an important line of research in psychometrics (Andrich & Master, 1988; Cox, 1980; Rojas, 2001).

Despite the widespread use of Likert-type scales in several fields of psychology and in spite of the various studies carried out, there is no consensus regarding the number of categories that a scale should possess. A number of authors (Alwin, 1992; Cicchetti, Showalter & Tyrer, 1985; Cox, 1980; McKelvie, 1978; Rodríguez, 2005; Wakita, Ueshima & Noguchi, 2012) have conducted exhaustive reviews of this type of study.

In general terms, in this field, scales are designed with different numbers of response options and the effect of this is observed in measurements of reliability and validity. However, the diversity of methodologies, measurements, and psychometric theories that have been applied make it difficult to draw clear conclusions. For example, in the analysis of reliability, the alpha coefficient (Cronbach, 1951) has been utilized as a measure of internal consistency. In these studies, instruments are applied that have between 7 and 25 answer choices. For the purposes of analysis, the number of options collapses and their effect is analyzed on the alpha coefficient. The results indicate that reliability of scales with homogenous items is barely affected by the number of response categories when compared with those whose items present greater heterogeneity (Mattell & Jacoby, 1971; McCallum, Keith & Wiebe, 1988; Weng, 2004).

On the other hand, reliability is researched as a measurement of stability by means of test-retest procedures (Boote, 1981; Chang, 1994; Weng, 2004). Unlike previous designs, these works showed that when the number of categories increases, reliability decreases. These studies also explore the effect of the assigned verbal labels of the various alternatives. In the results, an improvement is obtained in the reliability index when all response alternatives are labeled (Boote, 1981; Weng, 2004).

Four procedures have traditionally been used in validity studies: (a) analysis of the factorial structure, which has shown, in general, that the number of alternatives of the scale does not affect the validity of the instrument (Comrey & Montag, 1982; Vellicer & Stevensons, 1978); (b) correlation with other tests (Sancerini, Meliá & González-Romá, 1990); (c) evaluation of convergent and discriminating validity through multitrait-multimethod (MTMM) matrix analysis (Chang,1994), or (d) modeling by means of structural equations in which the results indicate that the model fits better to the measurement when the number of response alternatives increases (Ferrando, 2000) .

In recent decades, this issue has been explored within the framework of the Item Response Theory (IRT). For example, Hernández, Muñiz, & García-Cueto (2000) use the Graded Response Model (GRM) of Samejima (1969) and compare the number of cycles that the algorithm of estimation requires to achieve convergence when fewer or more alternatives are utilized. Ferrando (1999), on the other hand, analyzes three models: one of continuous variables; another for censored variables, and a multidimensional graded response model. Both studies suggest that adjustment improves as the number of alternatives in the scale increases to a limit of 6, as happens in the case of one-dimensional models. In more complex models, however, the increase does not produce an improvement in trait measurement.

In test theory, the characteristics of the population examined (age, schooling, ethnic group belonged to, etc.) are fundamental for test design and must be consider (Wakita, Ueshima & Noguchi, 2012). For example, it is a well-known fact that the capacity to discriminate increases with age; and therefore it is not appropriate to design tests with many alternatives for children. However, in the studies reviewed, the participants have been adults with similar characteristics regarding aspects such as educational attainment, social conditions, etc. In addition to these circumstances, some authors (Ferrando, 2000; Jöreskog, 1971) acknowledge the fact that most studies err in not proving the basic assumption that items with different formats are equivalent measurements of the same trait. For these authors, proving this supposition is not a mere formality, given that an item depends upon a variety of circumstances, among which format is one of the most important.

The majority of the research suggests that the number of response options is a factor that influences the reliability and validity of tests, although some studies have offered contradictory findings. These discrepancies can be explained by: 1) the psychometric model that is employed (classical test theory or item response theory); 2) the psychometric properties that are emphasized (validity or reliability) and 3) the method that is used to obtain the data (data collapsing, test-retest design). With these considerations in mind we believe that a contribution in the study of the number of response options should involve a change in the methodology that is used to obtain and analyze the data. To this end, we applied the 30-item revised scale that measures child stress perception in children, in which items were presented with 3, 5, and 7 response options. In the results, we examine 1) whether all items are measuring the same dimension, 2) the reliability of the various forms and 3) the convergent validity related to the instruments applied to parents and teachers. One can assume that the results will depend on the sample and will be different from those obtained with adult samples.

Methods

Participants

A total of 583 children between 10 and 12 years of age ($M_{age} = 11.3$) in the fourth and fifth grades of primary school participated in the study. Of these, 304 were girls and 279 were boys attending schools in the state of Michoacán, Mexico. Children at these schools participated in a two-week workshop about how to cope with drug violent situations in public places before they were tested.

Instruments

Stress perception scale: The 30 items in the revised Child Stress Perception Inventory (CSPI) were used to design 18 formats. Each format was constructed with three blocks of 10 items, each block having 3, 5, or 7 answer choices. Formats differed with respect to the items included in each block and the order of the blocks. Among the 18 formats, each item appeared 6 times in each block; the position of the items as well as the order of the blocks was randomly distributed. We also provided labels at the beginning and end of the response options the categories were: not nervous every option has an emoticon below that represent each state, each one differs in the shape of the mouth and in the number of the lines that stand for grades of agitation.

The format design permits the following: a) that all subjects respond to all options, and b) that all items are presented with all of the options.

Comparison List for Parents: We applied the Ackerman (1991) comparison list for parents. This is a twenty-three items instrument in which the parents respond with their assessment of certain behaviors and physical symptoms related with their children's anxiety. In the scale, the parents are asked if they have observed, in the past months, whether their son or daughter has displayed the behaviors that are indicated. The answer choices are: almost always; sometimes, or never.

Scale for teachers: The scale for teachers consists of three questions that ask the teacher to indicate to what degree he or she considers the child to be exposed to stressful situations and to what degree the child copes with these stressfull situations is able to confront these adequately. The questions and answer choices are the following:

1. The child has problems at home or at school.
2. The child is very affected by the problems he/she has.
3. The child is very nervous.

These latter three questions were answered on a scale of 7, in which 1 represents the least degree to which these behaviors were observed, and 7, the highest degree.

Procedure

The data analysis is divided into two phases: in the first, the possibility of establishing equivalency among nested models is investigated, while the second performs an analysis of the validity with which we analyze the results of the child stress test in relation to the scales for parents and teachers.

In the equivalency study, we attempt to prove the following assumptions: 1) that the items with a different number of response alternatives measure the same trait, and 2) that the parameter of location of the answer choices that share the same label among the distinct forms (3, 5, and 7 alternatives) are the same.

For treatment of data that derive from different formats, we propose a multidimensional model that comprises an extension of the Graded Response Model (GRM) of Samejima (1969). To prove assumption (1), we conducted a procedure of comparison of models between the multidimensional GRM and the one-dimensional GRM, and we appraised the loss of adjustment. This same procedure is employed to prove assumption (2). To this end, we restricted the parameters that correspond to the ordered categories in the GRM and observed whether there was an adjustment loss in the nested model with respect to the more general model. A likelihood ratio (LR) test was used to explore equivalency, in this test a restricted model in which the specified parameters are constrained to be equal is compared to a model in which those parameters are permitted to vary (augmented model). The test statistic is calculated as the difference in G^2 between the models, this statistic is distributed as a χ^2 with degrees of freedom equal to the difference in free parameters (*d.f(M)*).

Statistical significance indicates the loss of adjustment in the restricted model (Mood, Graybill y Boes, 1974).

A multidimensional formulation of the GRM: general model.

It has been said that in the case of comparing various response formats, there is a possibility that each item is measuring distinct traits; thus, we use for the general case a multidimensional extension of the GRM. In particular, we assume that the probability that person $j(j=1...n)$ responds to item $i(i=1...s)$ in category k or higher $(k=0...m_{ji}-1)$ (with *n* being the number of persons, *s* the number of items and m_{ji} the number of response categories in item *i* as presented to person *j*), is given by:

$$P^{*}(Y_{ji}\geq k)=\sum_{h=3,5,7}\frac{\exp[\alpha_i^{(h)}(\theta_i^{(h)}-\beta_{ik}^{(h)}]}{1+\exp[\alpha_i^{(h)}(\theta_i^{(h)}-\beta_{ik}^{(h)}]}I_h(j,i) \qquad [1]$$

$\alpha_i^{(h)}$ is the parameter of discrimination of item i when it is answered with h response categories. $\beta_{ik}^{(h)}$ is the parameter of difficulty or localization for category k of item i when the item is answered with h response categories. $\theta_j^{(h)}$ is the parameter of person j for the latent dimension that underlies the responses to the items that are answered with h response categories. $I_h(j,i)$ is the selector function that is defined as follows: $I_h(j,i)$ = 1 if item *i* was presented to person *j* with *h* response categories, and 0 otherwise.

To estimate the three-dimensional model, it is further assumed that the parameter of the person has been extracted from a trivariate normal distribution: $\theta_j^{(3)},\theta_j^{(5)},\theta_j^{(7)}\sim N(0,\Sigma)$, where the covariance matrix $\sum$ is the identity matrix.

Restrictions of the general model.

In Equation [1], in the case in which h is always the same (i.e., 3 answer choices), the model is reduced to GRM. In this regard, it has been stated that GRM is nested in the multidimensional GRM. The following restricted models were compared against the general model by a LR test:

$$H_0 : \theta_j^{(3)} = \theta_j^{(5)} = \theta_j^{(7)} = \theta_j \text{ for each } j \qquad [2.1]$$

This hypothesis signifies that the trait level of the person θ_j is the same when 3, 5, or 7 response categories are used, and it is evaluated on comparing the baseline multidimensional model with the one-dimensional GRM.

2.2.a) Ho: $\beta_{i2}^{(5)} = \beta_{i3}^{(7)}$; $\beta_{i4}^{(5)} = \beta_{i5}^{(7)}$ [2.2]
2.2.b) Ho: $\beta_{i1}^{(3)} = \beta_{i1}^{(5)} = \beta_{i1}^{(7)}$; $\beta_{i3}^{(3)} = \beta_{i5}^{(5)} = \beta_{i7}^{(7)}$; $\beta_{i2}^{(3)} = \beta_{i3}^{(5)} = \beta_{i4}^{(7)}$

Hypothesis 2.2 tests whether options with the same label in the formats with three, five, and seven alternatives have the same location parameter. Firstly, we test categories shared by the five and seven alternatives formats (2.2a), and then for the common categories to the three formats (2.2b). For example, in this study, the last category in all formats is the same ("extremely nervous"); meaning that these would have the same probability of response regardless of the item's response format ($\beta_{i3}^{(3)} = \beta_{i5}^{(5)} = \beta_{i7}^{(7)}$). Therefore, if we restrict the parameters that define the Response Function Categories (RFCs) that share labels and there is no loss in model fitting, one can conclude that adding answer choices does not alter the significance of the categories.

To establish the nested models that allow for equalizing these parameters and performance of the likelihood ratio test, we employ the property of univariance when adjacent categories are joined, which is exclusive of the GRM (Samejima, 1969). To prove the hypotheses, we used the Parscale 4.1 software program (Muraki & Bock, 2003). This program is designed for one-dimensional models; thus, we performed transformations in the data matrix in order to evaluate a multidimensional model. For the comparison, we followed the same logic with the one-dimensional model (see Table 1). In the table, an examinee is represented that responded to the first block of ten items with 3 alternatives, the second with 5, and the third block with 7 alternatives. In the first line, a sole subject is represented with 3 rows and 90 columns; in the first row, we find the responses to items presented with 3 options, in the second row those with 5 options, and in the third row, those with 7 options.

Items of those remaining are design-associated missing values (*missing* = 9). The collapsed multidimensional model possesses the same structure, except that the responses of 5 and 7 options are transformed into a scale of 3 options. In the case of strategy 2, the multidimensional model, each one examinee is represented by 3 rows with data transformed into 3 answer choices, that is, the same as in the previous case, but with only 30 columns, one for each item. For one-dimensional models, each one examinee is represented by a single row. This is carried out to prove whether conformation of the data does not affect the comparison's conclusions. Transformation was carried out first from 7 to 5 and then to 3 categories.

Table 1. Example of the data matrix ordering to execute the models comparison test

	3 options			**5 options**			**7 options**		
	block1	block2	block3	block1	block2	block3	block1	block2	block3
Multi-dimensional Model	**3**	9	9	9	9	9	9	9	9
	9	9	9	9	**5**	9	9	9	9
	9	9	9	9	9	9	9	9	**7**
Multi-dimensional Model (collapsed)	**3**	9	9	9	9	9	9	9	9
	9	9	9	9	**3**	9	9	9	9
	9	9	9	9	9	9	9	9	**3**
Estrategy 2	**3**	9	9						
Multi-dimensional Model	9	**3**	9						
	9	9	**3**						
One-dimensional Model	3	9	9	9	5	9	9	9	7
One-dimensional Model. (collapsed)	3	9	9	9	3	9	9	9	3
Estrategy 2 One-dimensional	3	3	3						

Results

A separate analysis for fitting the data matrices generated with 3, 5, and 7 answer choices to the GRM were conducted; PARSCALE offers a χ^2 of the all test that measures the adjustment of the test. A worst fit was obtained with 3 options (χ^2=276, *d.f.* = 295, *p* = 0.77), while best fit was achieved with 7 answer choices (χ^2=486, *d.f.* = 526, *p* = 0.89). For 5 options, we obtained (χ^2=374, *d.f.* = 406, *p* = 0.87).

To calculate the differences between the multidimensional and the nested unidimensional models, we utilized the Mplus 4.1 software program (Muthén & Muthén, 2006) and carried out the difference test in the goodness of fit statistic according to that proposed by Satorra (2000) for categorical variables. For the multidimensional model, we obtained ($G^2 = 26822.32, d.f. = 270$, correction factor = 0.861), and for the one-dimensional model, ($G^2 = 26825.69, d.f. = 273$, correction factor = 0.864). The difference between the models ($G^2 = 4.62$,*d.f.* = 3, $p > 0.05$) indicates that there is no significant loss of fit from a multidimensional to a unidimensional model, in other words, even when a different response format is employed, we are confronted with a unique trait, see hypothesis 2.1.

Table 2 presents the goodness of fit (G^2 statistic and degrees of freedom) for each model, and then the loglikelihood ratio test comparison (G^2 discrepancy between models, degrees of freedom difference ((*d.f.(M)*), and probability). The upper part of the table showed the test for the labels that are shared between 5 and 7 answer choices and among 3, 5, and 7 in the lower part. On comparing the general and the restricted models, the hypotheses that establish equivalency among the parameters of categories are rejected, these happened when a multi-dimensional or a unidimensional model was used regardless data matrix conformation (collapsing or using strategy 2), G^2 discrepancy in both strategies were similar, however there are more degrees of freedom when using strategy 2. As expected, the worst result is obtained when we fully collapse categories in a multi-dimensional model (χ^2=14463.5, *d.f(M)* = 80, p < 0.05).

Psychometric Properties

This section presents the analysis of the psychometric properties of the various response formats in relation to other variables. To this end, the following are demonstrated: the reliability obtained for each response format; the correlation between the trait value obtained through each of the formats; the correlation with the parents' check list and with the scale applied to the teachers. Lastly, the information function for each format is shown.

In Table 3, marginal reliability coefficients are presented in the diagonal for each of the response formats. As one can see, the highest reliability was obtained with 7 response-options scale (0.94). However, this index is only very different in the scales of 5 (0.93) or 3 answer choices (0.92). Alpha coefficients are presented for the check list which 378 parents ($\alpha = 0.96$) answered and also for teachers' questionnaire ($\alpha = 0.82$). Outside of the diagonal, the correlation is shown among the different forms and scales for parents and teachers; as can be observed, the highest correlations are obtained with the 5 answer-choices format.

Table 2. Models fit and the Loglikelihood Ratio Test (LRT) in nested models between the

augmented models (indicated with *) and restricted models.

Models	Models Fit		LR Test		
	G^2	*d.f.*	G^2 discrepancy	*d.f(M)*	p
7 y 5 Alternatives					
Multidimensional Model*	36096.90	360			
Multidimensional (collapsed)	32417.00	300	3679.90	60	<0.05
Estrategy 2 Multidimensional Model	32591.46	150	3505.43	210	<0.05
Unidimensional*	35477.43	360			
Unidimensional. (collapsed)	31831.87	300	3645.55	60	<0.05
Estrategy 2 Unidimensional	31995.99	150	3481.43	210	<0.05
3. 5 and 7 Alternatives					
Multidimensional Model*	47217.33	350			
Multidimensional Model (collapsed)	32753.82	270	14463.50	80	<0.05
Estrategy 2 Multidimensional Model	34539.16	90	12678.16	260	<0.05
Unidimensional*	45864.04	350			
Unidimensional (collapsed)	31539.86	270	14324.17	80	<0.05
Estrategy 2 Unidimensional	33476.94	90	12387.09	260	<0.05

Table 3. Alpha coefficient for each format (3, 5 and 7 Alternatives), parents check list and teachers questionnaire (on the diagonal). Correlation between the different scales (outside the diagonal).

	3 Al.	5 Al.	7 Al.	Parents	Teachers
3 Al.	0.92				
5 Al.	0.79(**)	0.93			
7 Al.	0.74(**)	0.75(**)	0.94		
Parents	0.60(**)	0.77(**)	0.71(**)	0.96	
Teachers	0.53(**)	0.68(**)	0.62(**)	0.67(**)	0.82

Information Functions

In Figure 1, the information function and the typical measurement error for each format are presented. As one can see, the 7 answer choice scale shows more information along all of the trait measurements, followed by that of 5 answer choices. There is a great difference between these latter two and the information function for 3 answer choices, which provides the lowest information levels. The latter, however, shows similar information values in high and very high trait levels, while in the 5- and 7 answer choice scales; these are centered and provide greater information in mean trait levels.

Figure 2 shows the relative efficiency function of the scales in the various formats. The efficiency function relative to the 7- with respect to the 5 answer choice scale is defined as:

$$ER(\theta;7,5)=\frac{I_7(\theta)}{I_5(\theta)} \tag{8}$$

and establishes that relative equivalence values = 1 indicate that both tests produced the same information at the considered trait level and that ER values >1 indicate that the 7-option test provides more information that that of 5 and whether 0 <ER <1 indicates that the 5-option test provides greater information at the level of the trait considered (Martínez-Arias, Hernández-Lloreda , & Hernández-Lloreda, 2006).

Figure 1. Information function (solid lines) and their corresponding standard error (dashed line) for each format.

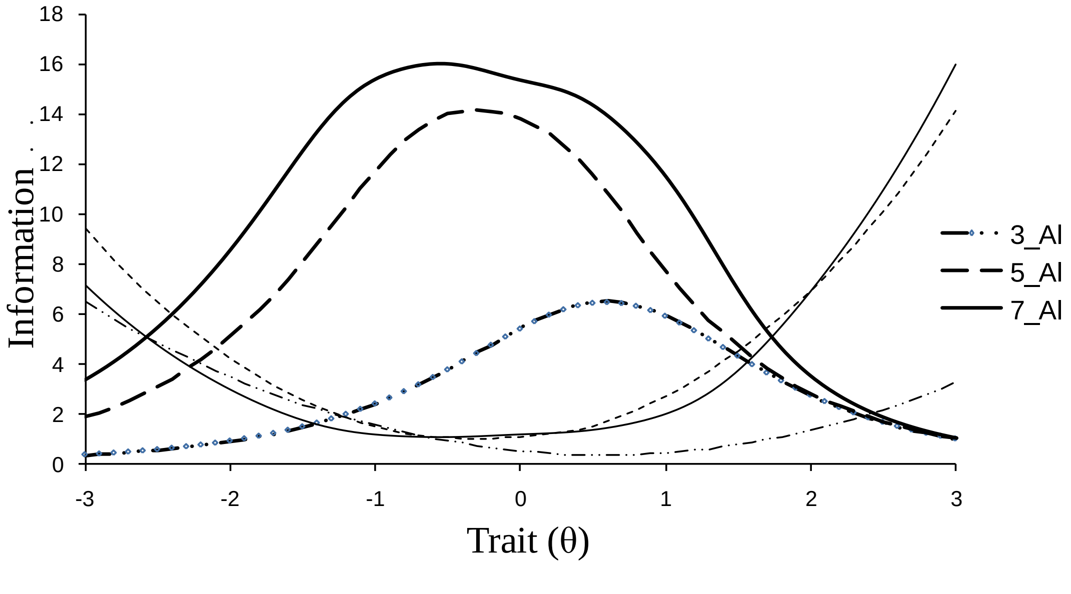
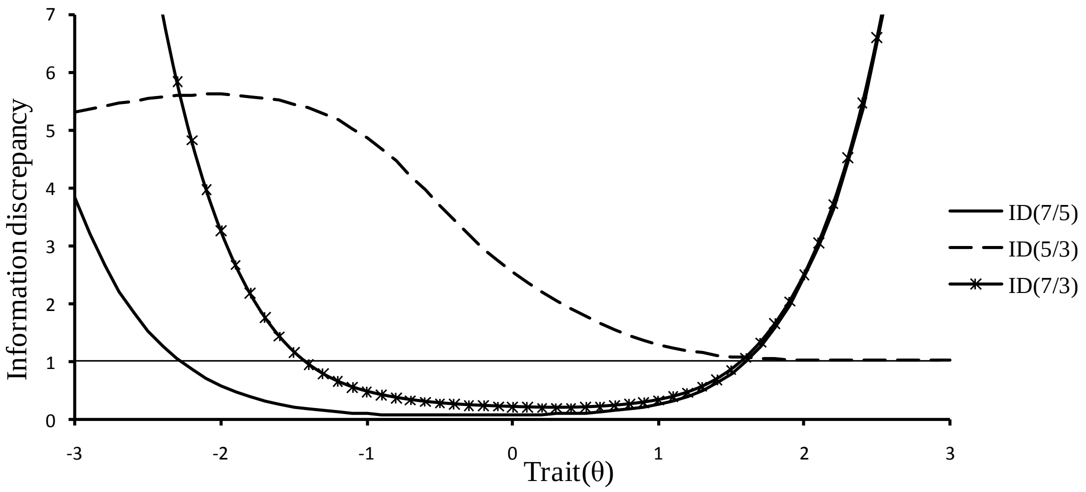


Figure 2 shows that the format with 7 answer choices is more informative than that of 5 answer choices only in low trait values; the same happens when we compare the 7-choice scale with respect to that of 3; the 5-choice scale only in low values of the trait; the same occurs if we compare the scale of 7 vs. that of 3. In the case of the 5-choice scale, this is better, or as informative, as that of 3 for all trait levels.

Figure 2. Function of relative efficiency of the test of 7 answer choices with respect to 5 (7/5), 5 answer choices with respect to 3 (5/3) and 7 with respect to 3 answer choices (7/3).



Conclusions

The development of a good psychometric test is the product of a process that entails various stages, including specifying the type of response format to use. In the field of personality evaluation in children, there is a tendency to employ dichotomous response formats or those with 3 answer choices as a maximum, based on the assumption that the lower discriminating capacity of the child negatively affects his/her attribution of judgment in scales of >3 points. Contrary to this assumption, the results of this study show that, at least in the instrument used here, the test has better psychometric properties with >3 answer choices. In general, the 5-alternative format proved to be better than the remaining two formats. The various empirical research projects or simulations designed to maximize the psychometric properties of the tests use a wide variety of strategies. In the case of this study, a novel strategy is presented that consists of allowing each subject to respond to a scale within the same session utilizing a different number of answer choices (3, 5, and 7). In our opinion, this strategy permits direct information gathering on the way in which the participants have really answered and avoids inferences derived from utilizing another type of methodology, such as collapsing data, with the subsequent loss of information and violation of the psychometric model employed (Ferrando, 2000), or comparison among different samples in the comparison designs among subjects. Given that design can affect the validity of the research, we have been particularly careful to prove two assumptions: 1) that the methodology would generate proportions of balanced data among the distinct formats, and 2) that the different conditions presented measure a sole trait.

Historically, studies on format began with methodologies based on the Classical Test Theory -CTT (Aiken, 1983; Bandalos & Enders, 1996; Cox, 1980), proceeding toward more sophisticated procedures from the viewpoint of measurement that derive from the Item Response Theory -IRT (such as García-Cueto, Muñiz, García-Cueto, & Lozano, 2005; Hernández, Muñiz & García-Cueto, 2000). At present, there is great consensus on the need to adjust the format of the characteristics of the construct measured, bearing in mind the characteristics of the target population of the psychometric application.

Together with the advance in the psychometric model, a change has been produced in the aspects that must be assembled in terms of the time required for establishing the effects of the format on the measurement of the construct. It is now known that reliability is not a sensitive indicator for establishing conclusions. More recently, validity analysis has been included as a substantive aspect (for example, Comrey & Montag, 1982; Olsson, 1979).

One of the main purposes of this study was to determine the collapse effect of the data, while being congruent with the labels shared by the categories. The results showed that data collapse can lead to a loss in model adjustment, which suggests that labels do not necessarily share the same significance if they are found in different formats. This result is important because it shows that despite the fact that collapsing categories is a very common practice, it is not always appropriate.

Regarding goodness-of-fit of the analysis (validity based on the internal structure) of the distinct single-factor models proposed for each scale in terms of the experimental conditions utilized (3, 5, and 7 options), an effect is observed on the number of answer choices on each variable. Thus, as the number of alternatives increases, goodness-of-fit improves. These results have also been reported by Ferrando (2000), Hernández, Muñiz & García-Cueto (2000) or García-Cueto, Muñiz, & Lozano (2002) and are contrary to the initial results carried out with CTT, which considers that the number of answer choices did not affect the internal structure of the scale (Mattell & Jacoby, 1971).

Moreover, the results obtained here indicate that the number of answer choices considerably affects the validity based on the relationship with other variables. Under the suppositions of the model, when we compare the correlation between the trait score under each experimental condition with its respective variables and criteria obtained from the parents and teachers, like Wakita, Ueshima & Noguchi (2012) we observe a significant effect of the number of answer choices on this relationship. However, these results do not correspond, in the case of children, to those obtained in adults in the studies conducted by Sancerini, Meliá, & González-Romá (1990), in that as the number of options increases, the criterion validity increases. In the case of the study with children, conditions change and validity shows better indices for 5, but this diminishes for 7 answer choices, although it is certainly better to utilize formats of 5 than formats of 3 answer choices. The latter conclusion is strongly reinforced by analyses of information functions, which demonstrate a significant loss of accuracy in employing the 3 answer choice format. The decision to use 5 or 7 points is more disputable, although we would bet on the 5-point format, which functioned better than that of 7 in terms of validity. The results of this research indicate the underlying need to analyze formats by means of equivalence studies. An additional possibility to that carried out here is the use of multiple indicator-multiple cause model (MIMIC) factorial analysis (Muthén, Kao, & Burstein, 1991), or Multitrait-multimethod (MTMM) matrices, which enable one to determine the conditions under which the response format can substantially affect the construct measured.

Final Note

This research was partly supported by a grant from the Universidad Nacional Autónoma de México (PAPIIT IN305810). The first author was supported by a doctoral fellowship from the PROMEP program of the Mexican Secretary of Public Education.

References

Ackerman, T. M. (1991). *Manual for the child behavior checklist/4-18 and 1991 profile*. Burlington: University of Vermont.

Aiken, L. R. (1983). Number of response categories and statistics on a teacher rating scale. *Educational and Psychological Measurement*, *43*, 397-401.

Alwin, D. (1992). Information transmission in the survey interview: number of response categories and the reliability of attitude measurement. *Sociological Methodology, 22,* 83-118.

Andrich, D., & Masters, G. (1988). Rating scales analysis. In J.P. Keeves (Eds.), *Educational research, methodology and measurement: an international handbook.* Elmsford, N.Y.; Pergamon Press.

Bandalos, D.L., & Enders, C. K. (1996). The effects of no normality and number of response categories on reliability. *Applied Measurement in Education, 9,* 151-160.

Boote, A. (1981). Reliability Testing of psychographic scales. *Journal of Advertising Research, 21,* 53-60.

Cicchetti, D.V., Showalter, D., & Tyrer, P.J. (1985). The effect of number of rating scale categories on levels of inter-rater reliability: A Monte-Carlo investigation. *Applied psychological Measurement, 9,* 31-36.

Chang, L. (1994). A psychometric evaluation of four-point and six-point likert type scales in relation to reliability and validity. *Applied Psychological Measurement, 18,* 205-215.

Cox, E. P. (1980) The optimal number of response alternatives for a scale: a Review. *Journal of Marketing Research, 17,* 407-422.

Comrey, A. L., & Montag, I. (1982). Comparison of factor analytic results with two choice and seven choice personality item formats. *Applied Psychological Measurement, 6,* 285-289.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of test. *Psychometrika, 16,* 297-334.

Ferrando, P.J. (1999). Likert scaling using continuous, censored, and graded response models: Effects on criterion-related validity. *Applied Psychological Measurement,* 23, 161-175.

Ferrando, P.J.(2000). Testing the equivalence among different item response formats in personality measurement: A structural equation modeling approach. *Structural Equation Modeling, 7,* 271-286.

García-Cueto, E., Muñiz, J., & Lozano, L,. M. (2002). Influencia del número de alternativas en las propiedades psicométricas de los test. *Metodología de las Ciencias del Comportamiento, volumen especial*.

Hernández, B. A; Muñiz, J., & García-Cueto, E. (2000). Comportamiento del modelo de respuesta graduada en función del número de categorías de la escala. *Psicothema*, 12, 288-291.

Jöreskog, 1971. Statistical analysis of sets of congeneric test. *Psychometrica,* 36, 109-133.

Martínez-Arias, R., Hernández-Lloreda, M.J., & Hernández-Lloreda, M.V (2006). *Psicometría*. Alianza Editorial: Madrid.

Matell, M. S. & Jacoby, J. (1971). Is there an optimal number of alternatives for likert scale items? Study I: reliability and validity. *Educational and Psychological Measurement,* 31, 657-674.

McKelvie, S. (1978). Graphic rating scales how many categories? *British Journal of psychology*, *69,* 185-202.

Mood, A., Gaybill, F. y Boes, D. (1974). Introduction to the theory of statistics. McGrawn-Hill International, London.

Muñiz, J., García-Cueto, E., & Lozano, L. (2005). Item format and the psychometric properties of the Eysenck Personality Questionnaire. *Personality and Individual Differences*, *38*, 61-69.

McCallum, D. M., Keith, B.R., & Wiebe, D. J. (1988). Comparison of response formats for multidimensional health locus of control scales: Six levels versus two levels. *Journal of Personality Assessment*, *52,* 732-736.

Muthen, B.O., Kao, Ch-F., & Burstein, L. (1991). Instructionally sensitive psychometrics: application of new IRT-based detection technique to mathematics achievement test items. *Journal of Educational Measurement*, *28,* 1-22.

Muthen, L.K., & Muthen, B.O. (2006). *Mplus: statistical analysis with latent variables: Users guide* (fifth edition). Los angeles, CA. Muthen & Muthen.

Muraki, E., & Bock, R.D.(2003). *Parscale 4.1*. Scientific Software International.

Rodriguez, M. (2005). Three options are optimal for multiple-choice items: a meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice,* 24, 3-13

Rojas, A. (2001). *Nuevos modelos para la medición de actitudes*. Promolibro, Valencia, pp. 214.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph Supplement*, 17.

Sancerini, M.D., Meliá, J.L., & González-Romá, V. (1990). Formato de respuesta, fiabilidad y validez, en la medición del conflicto de rol. *Psicologica*, *11*, 167-175.

Satorra, A., & P.M. Bentler (2001) A scaled difference chi-square test statistic for moment structure analysis , *Psychometrika*, 66,_507-514

Thissen, D. (1991). *MULTILOG user´s guide* (version 6.0) [computer manual]. Mooresville, IN: Scientific Sofware.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. En P. W. Holland y H. Wainer (Eds.). *Differential item functioning* (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum.

Vellicer, W.F., & Stevenson, J.F. (1978). The relation between item format and the structure of the Eysenck Personality Inventory. *Applied Psychological Measurement, 2,* 293-304.

Wakita, T., Ueshima, N., & Noguchi, H. (2012). Psychological distance between categories in the likert scale: comparing different numbers of options. *Educational and Psychological Measurement, 72,* 533-546.

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.