Content validity evidences in test development: an applied perspective

Delgado-Rico, Elena <javascript:contributorCitation('Delgado-Rico, Elena');>; Carretero-Dios, Hugo <javascript:contributorCitation('Carretero-Dios, Hugo');>; Ruch, Willibald <javascript:contributorCitation('Ruch, Willibald');>

Abstract: The purpose of this instrumental study was to show how to conduct a study aimed at obtaining content validity evidence in the test construction/adaptation process. An applied perspective was used, and this paper presents the content validity analysis of the Spanish adaptation of the State-Trait Cheerfulness Inventory trait form (STCI-T). This paper illustrates the stages required to analyze content validity: 1) definition of the content domain to be assessed, 2) item construction, and 3) expert judgment of the items constructed. This study focused mainly on the third stage and the results obtained with a previously selected panel of experts are included. The paper briefly describes the most important criteria to consider in the selection of experts, the procedure recommended to obtain judgments, the material to administer, aspects of items to assess, and the type of analyses that should be conducted. Based on the results obtained for the Spanish adaptation of the STCI-T, the article discusses the importance of obtaining content validity evidence in the test construction/adaptation process. The indices used demonstrated good content validity for the Spanish version of the STCI-T.

Originally published at:
Delgado-Rico, Elena; Carretero-Dios, Hugo; Ruch, Willibald (2012). Content validity evidences in test development: an applied perspective. International Journal of Clinical and Health Psychology España, 12(3):449-460.

© International Journal of Clinical and Health Psychology

ISSN 1697-2600 print
ISSN 2174-0852 online
2012, Vol. 12, Nº 3, pp. 449-460

Content validity evidences in test development: An applied perspective[1]

Elena Delgado-Rico *(University of Granada, Spain),*
Hugo Carretero-Dios[2] *(Universidad de Granada, Spain),* and
Willibald Ruch *(University of Zurich, Switzerland)*

ABSTRACT. The purpose of this instrumental study was to show how to conduct a study aimed at obtaining content validity evidence in the test construction/adaptation process. An applied perspective was used, and this paper presents the content validity analysis of the Spanish adaptation of the State-Trait Cheerfulness Inventory trait form (STCI-T). This paper illustrates the stages required to analyze content validity: 1) definition of the content domain to be assessed, 2) item construction, and 3) expert judgment of the items constructed. This study focused mainly on the third stage and the results obtained with a previously selected panel of experts are included. The paper briefly describes the most important criteria to consider in the selection of experts, the procedure recommended to obtain judgments, the material to administer, aspects of items to assess, and the type of analyses that should be conducted. Based on the results obtained for the Spanish adaptation of the STCI-T, the article discusses the importance of obtaining content validity evidence in the test construction/adaptation process. The indices used demonstrated good content validity for the Spanish version of the STCI-T.

KEYWORDS. Content validity. Test construction. Test adaptation. STCI-T. Instrumental study.

RESUMEN. El objetivo de este estudio instrumental es mostrar los pasos a seguir para la obtención de evidencias de validez de contenido dentro del proceso de construcción/

[1] This research was supported in part by the Department of Economy, Innovation and Science of the Andalusia Regional Government (Spain) under the Excellent Research Fund (SEJ-6569).

[2] Correspondence to: Hugo Carretero Dios. Facultad de Psicología. Universidad de Granada. Campus Cartuja s/n. 18071 Granada, Spain. E-mail: hugocd@ugr.es

adaptación de tests. Para ello se hace uso de una perspectiva aplicada, presentándose el estudio de validez de contenido llevado a cabo para la adaptación española de la versión rasgo del *State-Trait Cheerfulness Inventory* (STCI-T). Este trabajo profundiza en las fases que permiten obtener evidencias de validez de contenido: 1) definición de las áreas de contenido a evaluar, 2) construcción de ítems y 3) evaluación a través de expertos de los ítems construidos. Para este último punto se muestran los resultados encontrados para un panel de expertos previamente seleccionado. La presentación se centra en los criterios para la selección de expertos, procedimiento general a seguir, material para administrar, aspectos a evaluar de los ítems, y cálculos más importantes. Se termina argumentando sobre la relevancia de la validez de contenido en el proceso de construcción/adaptación de tests a partir de los resultados obtenidos para la adaptación española del STCI-T. Estos resultados ponen de manifiesto unos buenos índices de validez de contenido para los ítems de la versión española del STCI-T.

PALABRAS CLAVE. Validez de contenido. Construcción de tests. Adaptación de tests. STCI-T. Estudio instrumental.

From the first formal publication of the standards for test construction (American Psychological Association [APA], American Educational Research Association [AERA], and National Council on Measurement in Education [NCME], 1954/1999), there has been growing insistence on the need to provide evidence that the tests developed adequately represent the content domains of the construct assessed. More specifically, the various standards for test construction reflect the growing importance given to three concepts that are inseparable in the area of test construction: definition, representativeness, and relevance. These concepts formed the basic conceptual structure for the evaluation of content validity, which remains unchanged.

Although there are several definitions of content validity, most of them describe it as the degree to which the elements of an evaluation instrument are representative of the construct of interest (Haynes, Richard, and Kubany, 1995). For example, based on the need to thoroughly define the construct that is to be evaluated to assess this representativeness, Polit and Beck (2006) understand content validity as the extent to which an evaluation instrument contains an adequate sample of items for the construct assessed. Along the same lines, Wynd, Schmidt, and Schafer (2003) argued that content validity refers to the evidence needed to determine the degree to which an instrument adequately samples the research domain of interest. Content validity is therefore generally understood as the degree to which a sample of items represents an adequate operational definition of the construct of interest (Polit and Beck, 2006).

In addition to the item-domain conceptual relationship, the study of content validity comprises all the elements of items that directly affect the way data are obtained. Therefore, the formal aspects of items should be considered, since they also affect the way the construct is finally assessed on the scale (Haynes *et al*., 1995). Ambiguous or poorly drafted items, for example, do not fulfill the evaluation purpose because they yield biased responses, which implies not covering content-related aspects considered relevant for the construct assessed.

From these arguments, it is easy to conclude that content validity is an essential source of evidence and should be analyzed in any process of test construction/adaptation (APA, AERA, and NCME, 1999). Content validity evidence not only helps conceptually define the construct of interest but also lays the bases for a correct explanation of the variance in the scores obtained (Haynes *et al*., 1995). However, such evidence is rarely obtained and presented in detail, even though it would be highly desirable (Carretero-Dios and Pérez, 2007).

The objective of this instrumental study was to show the steps that should be considered in the process of obtaining content validity evidence in test construction/adaptation. The presentation is based on the key stages in the study of content validity: 1) definition of the content domain to be assessed, 2) item construction, and 3) expert judgment of the items constructed. The present study focused mainly on the third stage. Based on a traditional approach (Sireci, 1998) and an applied perspective, each of the stages mentioned is dealt with using the data obtained for the Spanish adaptation of the trait version of the *State-Trait Cheerfulness Inventory,* STCI-T (Ruch, Köhler, and van Thriel, 1996). The multidimensional nature of this instrument and the multifaceted approach to defining its dimensions make it ideal to clearly explain the steps that should be followed to obtain content validity evidence (see a review of the instrument in Ruch and Köhler, 2007).

Stages in the process of obtaining content validity evidence

1. Conceptual definition of the construct of interest

No content validity evidence can be obtained without specifically defining the construct to assess. The guarantees of the content validity evidence provided depend on how accurately the construct is defined and the extent to which the facets that delimit a construct are relevant for it (Haladyna, 2004).

An ambiguous definition or inadequate sampling of the construct components inevitably leads to poor representation of the construct in any scale eventually developed. Likewise, if no definition is explicitly provided, many of the results will be hard to interpret and it will be difficult to perform critical analyses of the instrument or its conceptual basis, which are so useful.

Conceptualizing the construct in the framework of obtaining content validity evidence requires starting by clearly defining its operational components and submitting the definition to expert judgment (Carretero-Dios and Pérez, 2007; Carretero-Dios, Pérez, and Buela-Casal, 2006). Several studies provide useful guidelines for delimiting the concept (Haynes *et al*., 1995) and preparing a formal definition of it (Osterlind, 1989). The example used in the present study, focused on the adaptation of the STCI-T, was dealt with following the theoretical proposal made by the authors of the original scale (for a more detailed analysis of the definition of the construct to assess, see Carretero-Dios *et al*., 2006).

The STCI-T was developed to assess the temperamental basis of sense of humor using the following three dimensions: 1) *Cheerfulness*, 2) *Seriousness*, and 3) *Bad mood*. Table 1 shows the definitions of these dimensions and the list of facets delimited for each of them (Ruch *et al*., 1996).

TABLE 1. Conceptual definition of the dimensions assessed with the trait version of the State-Trait Cheerfulness Inventory (STCI-T).

Cheerfulness (CH)	
Disposition or mood characterized by the usual presence of a feeling of enthusiasm, joy, joviality, etc. Cheerfulness is defined by the following components:	
CH_1	Prevalence of cheerful mood, bright and lively disposition.
CH_2	Low threshold for smiling and laughter, a tendency to laugh and express amusement very easily.
CH_3	Positive view of adverse life circumstances, a tendency to see the bright side of negative events and deal with them with optimism.
CH_4	View of a broad range of everyday stimuli as amusing, that is, a tendency to see the funny side of routine situations, searching for and enjoying anything that implies fun or cheerfulness.
CH_5	Generally cheerful interaction style, a preference for making people laugh or laugh with them and sharing celebrations or gatherings in which fun and laughter are present.
Bad Mood (BM)	
Disposition or mood characterized by the usual presence of a general feeling of affective discomfort, annoyance or resentment. Bad mood is defined by the following components:	
BM_1	Prevalence of a generalized state of bad mood, that is, general feelings of affective discomfort or displeasure.
BM_2	Prevalence of sadness (i.e., despondency, gloom).
BM_3	A difficulty enjoying oneself or showing joy, even in cheerful or funny situations.
BM_4	Prevalence of ill-humoredness (i.e., sullen and grumpy or grouchy feelings).
BM_5	Ill-humored behavior and attitudes in cheerfulness-evoking situations and toward such situations and the objects, persons, and roles involved.
Seriousness (SE)	
Attitude characterized by considering and facing most events and situations in life in a formal, grave and sober way. Seriousness is defined by the following components:	
SE_1	Prevalence of serious states (i.e., reflection, graveness, solemnity, formality, responsibility).
SE_2	Perception of even everyday happenings as important and tendency to consider them thoroughly and intensively.
SE_3	Tendency to plan ahead and set long-range goals and attaining a state as close as possible to personal well-being with the decisions and actions related to achievement of such goals.
SE_4	Tendency to prefer activities for which concrete, rational reasons can be produced and considering activities which don't have a specific goal or reason as a waste of time or nonsense.
SE_5	Preference for a sober, object-oriented communication style saying exactly what one means without exaggerating or ironic/sarcastic undertones.
SE_6	Rejection of cheerfulness-related behavior, roles, persons, stimuli, situations, and actions.

2. Item construction

In the present study, the 106 original items of the STCI-T (*Cheerfulness*: 38 items; *Bad mood*: 31 items; *Seriousness*: 37 items) were subjected to back-translation by four bilingual specialists (Hambleton and Jong, 2003). First, two specialists translated the items from the source language into Spanish; after that, the two other specialists

translated the items from Spanish back into the source language. Finally, the authors of the study and the translators discussed the results and developed a common proposal.

For the authors of this study, adapting a test should not be understood as merely translating the original items. Although the original items can be a useful anchor to begin the adaptation process, new items should be developed on the basis of the definition of the original construct. This leads to a better representation of the original construct for the new cultural context that is to be assessed. In an adaptation process, the basic reference is the definition of the construct that forms the basis of the test rather than a set of specific items developed for a given cultural context with peculiarities that may not be reflected in the new evaluation context.

Considering the definition of the construct, a new cluster of items was developed for each of the facets of the construct. After a joint discussion about the new items, the authors selected those that best represented the facets of the construct and met the guidelines for item-writing (Martínez, Moreno, Martín, and Trigo, 2009). This process led to an initial version of the STCI-T composed of 188 items (*Cheerfulness*: 66 items; *Bad mood*: 53 items; *Seriousness*: 69 items).

3. Expert judgment of items constructed

The basic objective of this stage is to analyze to what extent the items created are representative of the target construct and the degree to which such items represent the facet of the construct they were developed for, that is, their relevance (Beck and Gable, 2001; Mastaglia, Toye, and Kristjanson, 2003). As regards formal aspects, the classic criteria established by Angleitner, John, and Löhr (1986) were used as a reference. In this study, items were assessed on the basis of the following criteria: comprehension (assessment of whether the item is properly understood), ambiguity (judgment on the chances that the item can be interpreted in different ways), and clarity (extent to which the item is concise/accurate/direct).

A description of the study conducted to obtain expert judgment of the items developed for the Spanish version of the STCI-T is presented below. Each section includes detailed information on the study and contents that can be used to obtain further insight on this stage.

Method

Participants

Given the high number of items of the Spanish experimental version of the STCI-T (188) and the multifaceted nature of each of its dimensions, a large number of judges was selected. The aim was to divide the items to assess among the judges to avoid biases due to fatigue, loss of motivation in the task, or other causes. The number of experts selected was determined following the recommendations made by Crocker, Llabre, and Miller (1988) for obtaining useful estimates to adequately calculate interjudge agreement. The recommendation is to select at least three judges for each item (Lynn, 1986). However, along with this purely empirical criterion, the characteristics of the judges should be considered. Studies have highlighted the importance of involving

experts in test construction/adaptation and judges who are not experts in the measure but are specialized in the construct of interest or knowledgeable of the discipline it forms part of (Davis, 1992).

Combining the two previous criteria, a total of 18 judges were selected – 6 for each of the dimensions included in the STCI-T. The following experts were selected: nine teachers of the University of Granada with proven experience in test construction/ adaptation and nine PhD students from the Department of Social Psychology and Methodology of Behavioral Sciences of the University of Granada. To be eligible, judges had to have published recent scientific works on the subject.

Instruments

A booklet was prepared for each of the dimensions assessed by the STCI-T (*Cheerfulness, Bad mood,* and *Seriousness*). It included the instructions for the task, the conceptual definition of the construct and its facets, randomly arranged items, and aspects to assess for each item (representativeness, relevance, comprehension, ambiguity, and clarity*).* Representativeness, comprehension, ambiguity, and clarity were assessed using a 4-point Likert response scale (Davis, 1992). The relevance criterion included as many response options as the facets of the dimension. In this case, participants had to indicate which facet they considered each item corresponded to. In addition, each item included the possibility of proposing alternative wording (see an example of the data collection sheet in Appendix 1).

Procedure

After being invited to participate voluntarily, judges they were randomly given one of the three booklets and asked to complete them within one week. After completing the task, participants returned the booklet to the main researcher of the study. At that time, feedback was collected on their overall opinion about the task and the items, specific observations, and other relevant aspects.

Data analysis

The debate on the calculations that should be performed is still open (Beckstead, 2009; Landsheer and Boeije, 2010) and there are multiple approaches to the subject (Polit, Beck, and Owen, 2007; Wynd *et al*., 2003). Yet, the Content Validity Index (CVI; Polit and Beck, 2006) has traditionally been use to estimate representativeness, comprehension, ambiguity, and clarity. Although this index can be calculated in different ways, the authors of this study followed the recommendations made by Rubio, Berg-Weger, Tebb, Lee, and Rauch (2003). According to them, the CVI for each item should be calculated by dividing the number of judges issuing a judgment of 3 or 4 on the corresponding Likert scale by the total number of judges. The CVI was calculated for the relevance criterion by dividing the number of judges who considered that the item corresponded to the intended facet by the total number of judges. The CVI for the global dimension was calculated similarly after making the relevant decisions on the items. This was done by calculating the mean of the CVI for all the items conserved. As a general criterion, it is considered that CVI values should be $\geq .70$ (Tilden, Nelson,

and May, 1990). When there is a high number of items or the initial intention is to obtain clearly differentiated dimensions, a more restrictive criterion is recommended (Davis, 1992) with a minimum value of .80.

To analyze relevance, it is highly recommended to include an index of interjudge agreement that takes into account the number of judges and the number of classification possibilities as well as the total number of items when analyzing the global dimension. The recommendation is to use the interjudge agreement Kappa index (Wynd *et al*., 2003) with a value $\geq$ .40. The type of Kappa index used was that applied to categorical judgments made by multiple judges (Fleiss, 1971).

Decisions on items (*i.e.,* eliminating, modifying or conserving them) should not exclusively be based on empirical data. They should be subject to overall consideration by the authors depending on the objective intended when they were created, always based on the definition of the construct. It is also very useful to consider qualitative observations on items or alternative wording suggested for them.

Results

All the judges who were invited to assess the items completed the task. Of the 188 items assessed, 60 were considered to have insufficient content validity (CVI < .70 and Kappa < .40 in representativeness and/or relevance). In the dimensions of the STCI-T, this led to eliminating 16 items for *Cheerfulness*, 24 for *Seriousness*, and 20 for *Bad mood*.

The overall CVI value for representativeness was .89, .80, and .82 for *Cheerfulness*, *Seriousness* and *Bad mood*, respectively. As for relevance, overall CVI values were .81, .75, and .79 for *Cheerfulness*, *Seriousness*, and *Bad mood* respectively. The Kappa value was .55 for *Cheerfulness*, .48 for *Seriousness*, and .50 for *Bad mood*.

Table 2 shows a few results on representativeness and relevance. Information is provided only for some items taken as an example for the Bad mood dimension in its first facet, BM1 (see Table 1). This was done to illustrate the analyses conducted and the decisions made. This facet was chosen because it was the one that best illustrated the objective including only a few items.

TABLE 2. Example of results on representativeness and relevance.

			Representativeness	*Relevance*		*Action taken*
Dimension	*Facet*	*Examples of items*	*CVI*	*CVI*	*Kappa*	
Bad mood	BM 1 *Prevalence of a generalized state of bad mood*	People often have reason to ask is something is eating me	0	.16	.07	Eliminated
		My mood is often not the best one	.33	.16	.07	Eliminated

TABLE 2. Example of results on representativeness and relevance. (Cont.)

			Representativeness	*Relevance*		*Action taken*
Dimension	*Facet*	*Examples of items*	*CVI*	*CVI*	*Kappa*	
		I am often in a bad mood	.66	.66	.4	Kept
		If I am in a bad mood, I can't stand the presence of cheerful people	1	0	0	Eliminated
		When I am in a bad mood, I tend to be less considerate of others	.83	0	0	Eliminated
		There are many days on which I think, "I should have stayed in bed"	.66	.40	.30	Kept after Modification: Because of my gloomy mood, there are many days on which I think, "I should have stayed in bed"
		I could be described as a person with a "damaged" mood	.33	.40	.20	Eliminated
		My mood is usually bad	.66	.66	.40	Kept
		I often say to myself I didn't have a good day	.66	.66	.40	Kept

Note: CVI = Content Validity Index; Kappa = Interjudge agreement Kappa index.

All the items kept after the study of their representativeness and relevance showed adequate values for the indices of comprehension, ambiguity, and clarity (CVI $\geq$ .70). The following global values were obtained for each dimension: *Cheerfulness* (CVI comprehension = .95; CVI ambiguity = .92; CVI clarity = .92), *Seriousness* (CVI

comprehension = .92; CVI ambiguity = .92; CVI clarity = .80), and *Bad mood* (CVI comprehension = .95; CVI ambiguity = .80; CVI clarity = .80).

Discussion

The present study shows the steps that should be taken into account to obtain content validity evidence. The example provided was taken from an adaptation process instead of a construction process. Yet, adaptation processes also require reviewing the theoretical proposal of the original scale and submitting all the relevant information on the definition of the construct and the items clearly and accurately. For this reason, the information presented is also applicable to the construction process.

The data obtained show that the process of obtaining content validity evidence leads to an improvement of the items created both regarding the formal wording aspects and the theoretical representativeness-relevance of such items. Thus, obtaining content validity evidence makes it possible from the outset to provide empirical data supporting the construction/adaptation process (Sireci, 1998), which also facilitates the subsequent stages (Carretero-Dios, Pérez, and Buela-Casal, 2009).

As mentioned above, the adaptation process should not be understood as a mere translation of the original items to be validated in a new context. The authors of this study consider that the adaptation process would be enriched by creating new items based on the definition of the original construct. This would not only provide greater guarantees of obtaining an appropriate adaptation to the new evaluation context but would also broaden the view of validity studies by considering various samples, cultures and items at the same time.

Finally, the adaptation process presented here was conducted from a traditional approach to obtain content validity. Traditional procedures have some limitations that have led certain authors to suggest using multidimensional scaling procedures (Sireci, 1998). Some of the limitations mentioned are, for example, social desirability in expert judgments or the tendency to obtain medium or high values when judging the items. However, it has been argued (Rubio *et al*., 2003) that the key to overcoming the limitations of the traditional approach is to make an appropriate and representative selection of experts that guarantees their thorough assessment. It should also be noted that empirical analyses do not preclude a thorough analysis of the responses of experts and that the final decisions on the items should not rest on a specific analysis.

It would be interesting to conduct studies aimed at comparing traditional procedures and procedures based on multidimensional scaling. This would provide a good contribution for the study area of test construction/adaptation and would have important applications for areas of evaluation of growing social interest (*e.g.,* Aguayo, Vargas, Fuente, and Lozano, 2011; Escarpín, Rodríguez-Carballeira, Gómez-Benito, and Zapf, 2010; Fonseca-Pedrero, Sierra-Baigrie, Paino, Lemos-Giráldez, and Muñiz, 2011; Goñi, Madariaga, Axpe, and Goñi, 2011; Ortet *et al*., 2010; Verdugo, Arias, Gómez, and Schalock, 2010).

References

Aguayo, R., Vargas, C., Fuente, E.I., and Lozano, L.M. (2011). A meta-analytic reliability generalization study of the Maslach Burnout Inventory. *International Journal of Clinical and Health Psychology, 11,* 343-361.

American Psychological Association [APA], American Educational Research Association [AERA], and National Council on Measurement in Education [NCME] (1954, 1999). *Standards for educational and psychological tests and manuals*. Washington, DC: American Psychological Association.

Angleitner, A., John, O.P., and Löhr, F.J. (1986). It's what you ask and how you ask it: An itemmetric analysis of personality questionnaires. In A. Angleitner and J. S. Wiggins (Eds.), *Personality assessment via questionnaires. Current issues in theory and measurement* (pp. 61-108). Berlin, Germany: Springer.

Beck, C.T. and Gable, R.K. (2001). Ensuring content validity: An illustration of the process. *Journal of Nursing Measurement, 9,* 201-215.

Beckstead, J.W. (2009). Content validity is naught. *International Journal of Nursing Studies, 46,* 1274-1283.

Carretero-Dios, H. and Pérez, C. (2007). Standards for the development and review of instrumental studies: Considerations about test selection in psychological research. *International Journal of Clinical and Health Psychology, 7*, 863-882.

Carretero-Dios, H., Pérez, C., and Buela-Casal, G. (2006). Dimensiones de la apreciación del humor. *Psicothema, 18,* 465-470.

Carretero-Dios, H., Pérez, C., and Buela-Casal, G. (2009). Content validity and metric properties of a pool of items developed to assess humor appreciation. *Spanish Journal of Psychology, 12,* 773-787.

Crocker, L., Llabre, M., and Miller, M.D. (1988). The Generalizability of Content Validity Ratings. *Journal of Educational Measurement, 25*, 287-299.

Davis L.L. (1992). Instrument review: Getting the most from a panel of experts. *Applied Nursing Research, 5,* 194-197.

Escarpín, J., Rodríguez-Carballeira, A., Gómez-Benito, J., and Zapf, D. (2010). Development and validation of the workplace bullying scale EAPA-T. *International Journal of Clinical and Health Psychology, 10,* 519-539.

Fleiss, J. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76,* 378–382.

Fonseca-Pedrero, E., Sierra-Baigrie, S., Paino, M., Lemos-Giráldez, S., and Muñiz, J. (2011). Factorial structure and measurement invariance of the Bulimic Investigatory Test, Edinburgh across gender and age. *International Journal of Clinical and Health Psychology, 11,* 109-123.

Goñi, E., Madariaga, J., Axpe, I., and Goñi, A. (2011). Structure of the Personal Self-Concept (PSC) Questionnaire. *International Journal of Clinical and Health Psychology, 11*, 509-522.

Haladyna, T. (2004). *Developing and validating multiple-choice test items*. New York: Lawrence Erlbaum Associates.

Hambleton, R.K. and Jong, J.H. (2003). Advances in translating and adapting educational and psychological tests. *Language Testing, 20,* 127-134.

Haynes, S.N., Richard, D.C.S., and Kubany, E.S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7,* 238-247.

Landsheer, J.A. and Boeije, H.R. (2010). In search of content validity: Facet analysis as a qualitative method to improve questionnaire design. *Qualitity and Quantity, 44,* 59-69.

Lynn M.R. (1986). Determination and quantification of content validity. *Nursing Research, 35,* 382-385.

Mastaglia, B., Toye, C., and Kristjanson, L.J. (2003). Ensuring content validity in instrument development: Challenges and innovative approaches. *Contemporary Nurse, 14,* 281-291.

Martínez, R.J., Moreno, R., Martín, I., and Trigo M.E. (2009). Evaluation of five guidelines for option development in multiple-choice item-writing. *Psicothema, 21,* 326-330.

Osterlind, S.J. (1989). *Constructing test items.* London: Kluwer Academic Publishers.

Ortet, G., Escrivá, P., Ibáñez, M., Moya, J., Villa, H., Mezquita, J., and Ruipérez, M. (2010). Versión corta de la adaptación española para adolescentes del NEO-PI-R (JS NEO-S). *International Journal of Clinical and Health Psychology, 10,* 327-344.

Polit, D.F., and Beck, C.T. (2006). The content validity index: are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health, 29,* 489-497.

Polit, D.F., Beck., C.T., and Owen, S. (2007). Is the CVI an acceptable indicator of content validity? *Research in Nursing & Health, 30,* 459-467.

Rubio, D.M., Berg-Weger, M., Tebb, S.S., Lee, E.S., and Rauch, S. (2003). Objectifying content validity: Conducting a content validity study in social work research. *Social Work Research, 27,* 94-104.

Ruch, W. and Köhler, G. (2007). A temperament approach to humor. In W. Ruch (Ed.), *The sense of humor: Explorations of a personality characteristic* 2nd ed. (pp. 203-230). Berlin: Mouton de Gruyter.

Ruch, W., Köhler, G., and van Thriel, C. (1996). Assessing the «humorous temperament»: Construction of the facet and standard trait forms of the State-Trait-Cheerfulness-Inventory—STCI. *Humor: International Journal of Humor Research, 9,* 303–339.

Sireci, S.G. (1998). Gathering and analyzing content validity data. *Educational Measurement, 5,* 299-321.

Tilden V.P., Nelson C.A., and May B.A. (1990). Use of qualitative methods to enhance content validity. *Nursing Research, 39,* 172-175.

Verdugo, M.A., Arias, B., Gómez, L., and Schalock, R. (2010). Development of an objective instrument to assess quality of life in social services: Reliability and validity in Spain. *International Journal of Clinical and Health Psychology, 10,* 105-123.

Wynd, C.A., Schmidt, B., and Schaefer, M.A. (2003). Two quantitative approaches for estimating instrument content validity. *Western Journal of Nursing Research, 25,* 508-518.

Received October 12, 2011
Accepted February 2, 2012