Soc Indic Res (2013) 114:963–975
DOI 10.1007/s11205-012-0184-8

How Cognitive Interviewing can Provide Validity Evidence of the Response Processes to Scale Items

Miguel Castillo-Díaz · José-Luis Padilla

Accepted: 24 October 2012 / Published online: 3 November 2012

Abstract The current theory about validity reflected in the *Standards for Educational and Psychological Testing* (AERA et al. in Standards for educational and psychological testing, American Psychological Association, Washington, DC, 1999), offers no clear indications about the methods for gathering validity evidence about the response processes. Cognitive interviewing (CI) can play an important role answering the current demand about empirical and theoretical analyses of the response processes as a source of validity evidence in psychological testing. CI can provide validity evidence for investigating substantive aspects of construct validity and for contributing to the explanations for item and test scores (Zumbo in Handbook of statistics, vol 26, Elsevier, Amsterdam, pp. 45–79, 2007; The concept of validity: revisions, new directions and applications, IAP—Information Age Publishing Inc., Charlotte, NC, pp. 65–82, 2009). The aim of the study was to illustrate the use of cognitive interviewing method for gathering validity evidence on response processes. The search for evidence about the "response process" was guided by an argument-based approach to validity (Kane in Psychological Bulletin 1992; Educational measurement, American Council on Education/Praeger, Washington, DC, pp. 17–64, 2006). 21 cognitive interviews were carried out during the cognitive testing of the APGAR psychological scale intended to measure the "family support" construct. Cognitive interviewing provided validity evidence that explains how respondents interpret and respond to the APGAR items. Respondents maintained a shared interpretation of "family concept" while answering the APGAR scale items. Nevertheless, they included in the concept of family not only family members they live with but also other family members and even friends. CI participants were also capable of classifying their answers about the family support perception following a polythomous response system. Lastly, the role of CI in the Kane's argument-based approach and Zumbo's contextualized view of validity will be discussed.

Keywords Cognitive interviewing · Construct validity · APGAR · Cognitive response processes

M. Castillo-Díaz (✉) · J.-L. Padilla
Department of Methodology of Behavioral Sciences, School of Psychology, University of Granada, Campus de Cartuja, 18071 Granada, Spain
e-mail: miguelcastillo@ugr.es

1 Introduction

In recent years there has been a broad consensus on the pivotal role of validity in the development and evaluation of psychological tests and scales. However, there are open discussions about the significance of the validity or whether the consequences of using of the tests are or are not an issue of validity (Messick 1989; Sijtsma 2009; Lissitz 2009). A less discussed and equally unresolved issue is that of the methods to obtain evidence of validity and how to interpret, especially in the case of sources of validity evidence not included in previous editions of the *Standards for Educational and Psychological Testing* (AERA, APA and NCME 1999).

The latest edition available of the *Standards* (AERA et al. 1999) proposes a framework for validation studies based on five "sources of validity evidence". One source of validity evidence incorporated in this edition is that of the "evidence based on the response processes". According to the *Standards* (AERA et al. 1999), evidence based on response process refers to "evidence concerning the fit between the construct and the detailed nature of performance or response actually engaged in by examinees" (p. 12). As noted by Sireci (2009), among the few indications for obtaining evidence about the response processes, the following are mentioned "interviewing test takers about their responses to test questions, systematic observations of test response behaviour, evaluation of the criteria used by judges when scoring performance tasks and analysis of item response time data" (p. 18).

Currently, few studies have addressed the collection of validity evidence about the response processes (DeWalt et al. 2007; Ding et al. 2009; Gadermann et al. 2011; Ercikan et al. 2010; Irwin et al. 2009; Olt et al. 2010; Poole et al. 2009; Willis et al. 1991). In fact, as shown by Zumbo and Shear (2011) in the last 20 years the number of publications about the search for evidence of response processes has hardly increased, while there is a wide disparity in comparison with other "traditional" sources of validity evidence. In most of the cases, cognitive interviewing or a similar method was applied to investigate the response processes. DeWalt et al. (2007) used cognitive interviews to evaluate ambiguity, and understand how respondents see the relevance of a set of psychological items. Ding et al. (2009) conducted interviews to identify and resolve validity issues that stem from the differing perspectives of students and revealed validity issues missed by experts. Gadermann et al. (2011) conducted think aloud protocol interviews to examine the cognitive processes of children when responding to the items of the Satisfaction with Life Scale. They found that most of the children's responses were based on either an absolute strategy to indicate the presence or absence of something that is important for their judgments of their satisfaction, or a relative strategy using comparative statements. Ercikan et al. (2010) applied a think aloud method to confirm sources of differential item functioning (DIF). They focused on examining the extent to which linguistic differences identified by expert reviewers in the previous research were supported by evidence from the think aloud protocols. Irwin et al. (2009) resorted to cognitive interviewing procedures to gain feedback on response processes to items measuring physical functioning, emotional health, social health, fatigue, pain and asthma-specific symptoms. Poole et al. (2009) evaluated the NePiQoL scale of the "quality of life" construct through cognitive interviewing to identify items with redundant content and obtain evidence of validity. Meanwhile, Olt et al. (2010) interviewed fifteen respondents to investigate their response processes as a way to collect evidence on test validity; they carried out the testing of the response processes to identify problems in relation to respondents' conceptualization of cultural competence in the psychometric evaluation of a Swedish test instrument. Willis et al. (1991) elicited verbal reports from respondents to uncover and resolve validity issues in two health survey questionnaires.

On the other hand, Sijtsma (2009) reflected the general feeling of frustration among researchers in the absence of clear information on how to conduct validation studies, specifying that "in principle this methodology implies an endless array of research projects, which threatens to result in indecisiveness with respect to construct validity" (p. 168). This lack of guidance, adds to the absence of a general outline on how to interpret the evidence and what to get it. For example, Zumbo (2009) noted that we have focused overly heavily on the validation process and as a result we have lost our way and he emphasizes the need to re-focus our attention on why we are conducting all of these analyses.

With this situation in mind Michael T. Kane, in 1992, proposed an argument-based approach to validity, offering a way out of this cul-de-sac. Kane (1992) encouraged professionals to think of validation in terms of 'validity arguments' rather than in terms of 'research validity.' According to Sireci (2005): "Kane's practical perspective is congruent with the current *Standards*, which provide detailed guidance regarding the types of evidence that should be brought forward to support the use of a test for a particular purpose" (p. 3).

Kane (2006) states that to establish the "validity argument" it is necessary to begin by clarifying the content of the proposed interpretation of the measurements, via the "interpretive arguments". To construct the interpretive argument, first, one must specify the scheme of inferences and assumptions leading from scores on the test to the conclusions and decisions, and, secondly, conduct specific studies that support the plausibility of the proposed interpretations. The more rigorous the development of an interpretive argument, the clearer the content of the proposed interpretation of the measurements (Padilla et al. 2007). The interpretive argument orientates the validation towards indicating the type of validity evidence to study, with a view to examining the assumptions that support the desired interpretation of the scores (Kane 1992).

Chapelle et al. (2010) demonstrated that Kane's perspective offers new insights into validation, and achieves the aim of making validation more accessible, developing a validity argument for the proposed interpretations and uses of the *Test of English as a Foreign Language* (TOEFL). Likewise, Llosa (2008) used an argument approach to validate the quality of teacher judgments in the context of a standards-based classroom assessment of English proficiency, affirming that this approach illustrates how to identify and organize the collection of relevant supporting evidence. Both studies are examples of Kane's (2010) proposal: "the validation process requires many different kinds of analysis and evidence, some of the evidence may be empirical (e.g., studies of reliability, equating, relationships to other variables) and some will be judgmental (analyses of how scoring keys were developed and applied), and the validity argument relies on all of this evidence to reach a general conclusions about how much confidence we can place in the proposed interpretations and uses" (p. 181).

As noted earlier, the cognitive interview method or a similar method has been used to obtain evidence related to the response processes (DeWalt et al. 2007; Ercikan et al. 2010, Irwin et al. 2009; Poole et al. 2009; Willis et al. 1991). The logic behind this study is that the cognitive interview can answer the current demand of theoretical and empirical analyses of the response processes, and it can be applied to gather validity evidence to support the use of a questionnaire in an argument-based approach framework. This rationale is in line with the strong argument made by Gadermann et al. (2011), in favour of CI as a method for investigating the processes that are involved in the measurement task. Such research is aimed at obtaining validity evidence on the substantive aspects of construct validity (Messick 1995).

DeVellis (2003) notes that similar to input from expert researchers and clinicians, cognitive interviewing contributes to both the validity and reliability of measures by providing data on the relevance and clarity of questionnaire items. The growing use of cognitive interviews points to the importance of qualitative data in assessing the validity of measurements obtained by tests and questionnaires (Castillo et al. 2010; Knaff et al. 2007).

The cognitive interview is one of the predominant methods for identifying and correcting problems with survey questions (Beatty and Willis 2007). Cognitive interviewing is mainly used in survey pretesting, when survey researchers try to identify systematic errors in the "question-and-answer" process (Willis 2005). Based mainly on a cognitive four-stage model, cognitive interviewing explores the various stages of the "question-and-answer" process, these are: interpretation of the question, information retrieval, judgment and communication of the response (Tourangeau 1984). It provides evidence about problems with the comprehension of key terms, failures in the data retrieving, errors in the wording of the question and mismatch in the choice of response options designed for the questionnaire. Cognitive interviewing is a powerful method for understanding the thought processes of the respondents when answering questions (Beatty and Willis 2007). In addition, this method is often used to explore the response processes of behavioural frequency questions (e.g., Menon 1993; Bickart and Felcher 1996; Conrad et al. 1998). Although the application of cognitive interviewing has been mainly related to questionnaires included in surveys, it can also be implemented for assessing test and scale items. The usefulness of the cognitive interview depends not so much on the type of question to evaluate, as the processing that the participant performs. The method is useful when processing is not automatic, that is, when respondents have access to their thinking, when the respondent may recall and count specific events, storing a description of the events and a running tally in working memory (Conrad and Blair 2009). Whereas, if the processing were automatic, the participant would have little or nothing to say about how he got the response and the cognitive interview would be meaningless (Conrad and Blair 2009).

The aim of this study is to illustrate the use of cognitive interviewing method to gather validity evidence on response processes. The validation study is carried out using the argument based approach, with the aim of showing how the cognitive interview method may be useful to obtain evidence of validity based on the intended use of scores. This study was performed to validate the APGAR scale on family function. This scale was developed by Smilkestein in 1978 to assess the construct: family support. The APGAR scale has been frequently used by official statistical institutes in health surveys. To report cognitive interviewing findings, this study follows the Cognitive Interviewing Reporting Framework (CIRF) proposed by Boeije and Willis (2011) as a quality framework that cognitive interviewers can use when reporting cognitive interviewing studies.

2 Method

2.1 Participant Selection

The cognitive interviews were conducted with 10 men and 11 women, aged between 20 and 67 years (seven participants aged between 20 and 35 years old, seven participants aged between 36 and 50, and seven participants between 51 and 67 years old). The variables that guided the selection of participants were: (a) equal gender distribution, (b) age range 18–70 years old; (c) variety in marital status ("marital status"): seven single, 10 married, three divorced and two widowed, (d) balanced distribution of educational level: eight participants with a basic level, seven with an intermediate level and six with a higher level of studies, (e) type of housing: nine respondents who lived alone (single households) and 12 people living with other family members. An intentional selection was conducted, recruiting people whose profile could provide evidence to prove the interpretive argument, the subject of the validation study. This argument would be based on the estimated scale of the level of satisfaction with family support among household members. For instance, participants with different marital status and who were living in different types of household were recruited, trying to capture different interpretations of the "family function" construct. The participants were recruited via "mouth-of-mouth".

2.2 Materials

2.2.1 The APGAR Scale

The APGAR scale designed by Smilkstein (1978) is used in the clinical practices of family physicians as a tool to get information quickly and easily on the family situation and its possible role in the origin and resolution of conflicts (Bellón et al. 1996). The APGAR scale evaluates five components of family function using five Likert items: *adaptability, partnership, growth, affection* and *resolve* (Bellón et al. 1996). Spanish adaptation was performed using a back translation design. The Cronbach Alpha value of the Spanish version of APGAR scale responses in the studies used is around .84, also the factor analyzes performed supported the unidimensionality of the scale (Bellón et al. 1996). Table 1 presents the original version of the APGAR scale.

In addition, measures provided by APGAR have been related to "frequency of medical visits", "immunological responses", "emotional stress" and "depressive symptoms" (Smilkstein et al. 1982).

The APGAR scale was included in the adult questionnaire of the Spanish National Health Survey (SNHS) (Spanish Ministry of Health and Consumption 2006).

Table 1 Original APGAR Scale

Items	Almost always	Some of the time	Hardly ever
1. Are you satisfied with the help you receive from your family when you have a problem?	□	□	□
2. Are you satisfied with the time you and your family spend together?	□	□	□
3. Do you feel you family loves you?	□	□	□
4. Do you talk together about problems you have in home?	□	□	□
5. Important decisions are made by all of you together in home?	□	□	□

2.3 Ethics and Data Collection

First, participants were informed about the purpose of the study. To motivate participants, they were told how important the interviews will be to improve a national health survey promoted by an official body. The interviewers told also the participants the information provided by the survey will be used by policy makers. In addition, each participant was rewarded after finishing the interview with 30 Euros. The interviews were conducted individually by four trained and experienced interviewers (three females and one male). The interviews were recorded on audio and video with the consent of the participants and took place in a cognitive laboratory. The participants were guaranteed confidentiality and that the data would be used solely for purposes related to research.

2.4 Interpretive Argument

In accordance with the approach to validity based on arguments (Kane 2006), an interpretive argument was developed to guide the validation study. It is important to note that the application method for the adult questionnaire of the SNHS was personal interviewing in pre-selected homes. In this way, the context of the application of the SNHS was: (a) the households were the sampling units of the SNHS, and (b) the aim was to estimate the level of satisfaction with "family support" in the Spanish population coming from the responses given by the person in the home who meets the survey requirements. Table 2 shows the intended use of the APGAR scale scores, the interpretive argument and the supposed purpose of the validation study. Following Kane's indications (Kane 2006, 2010), assumptions were considered to be key to supporting the interpretative argument taking the methodological context and the application method for the SNHS.

2.5 Research Design

Cognitive interviewing of the APGAR scale was performed for the pretest of the SNHS (Spanish Ministry of Health and Consumption 2006).To carry out the cognitive interviews, a "probing based" paradigm was applied (Beatty and Willis 2007) which included general and specific probes. The paradigm is based on the well-known "question-and-answer" proposed by Tourangeau et al. (2000). Interviewers used an interview protocol which included follow-up probes. Cognitive interviewing protocol is characterized by the fact that the questions and follow-up probes guide the interaction, also giving the interviewer the freedom to explore relevant issues. The follow-up probes were applied retrospectively, that

Table 2 Aim/content of the validation study

Intended use for APGAR: To estimate the level of satisfaction with "family support" of Spaniards from sample data provided by SNHS

Interpretative argument: APGAR scores are measures of "family support" construct

Assumption 1: Respondents answer the APGAR scale items interpreting the construct "family support" and keeping in mind a constant meaning for the "family concept" across APGAR items

Assumption 2: Given that the APGAR questionnaire is included in a survey whose sample unit is the household, interviewees should respond to APGAR items taking mainly into account people who live in their homes when the questionnaire is applied

Assumption 3: Given the APGAR questionnaire is composed of 5 multiple choice items, respondents are able to grade their response and adjust it to multiple choice options

Table 3 Outline for the analysis of cognitive interviews

Interpretive argument	Probes
Assumption 1: Meaning of "family concept"	"What were you thinking about when answering the questionnaire?" "How have you understood *recieving help from the family when you have a problem*?"
Assumption 2: Family members	"How many people make up your close family?" "Who in your family helps you when you have a problem?"
Assumption 3: Responding by a polythomous system	"In the question, *Do you feel that your family loves you? You answered* '…'. In what situation would you answer '(higher or lower alternative to the answer given)'?"

is, first the APGAR scale was administered, and then the follow-up probes. The retrospective application of the probes is appropriate when the presentation of the items is desired to be as realistic as possible (Willis 2005).

The content of the probes was developed through an expert opinion procedure. The tests should provide evidence to examine the assumptions that underpin the interpretive argument. Table 3 shows examples of the tests along with the assumptions.

2.6 Data Analysis

Video files of the interviews were inputted in the program AQUAD, version 6.8.1.1 (Huber 2008) for analysis. This program allowed the identification of the response patterns of the respondents using codes and annotations in the video files of the interviews. For each of the probes a code was assigned reflecting whether the participant's response process included the interpretation provided in the interpretive argument, and if not, what interpretation took place. By this assignment of codes to those portions of interviews the evidence on t he response processes of respondents could be extracted. The analysis of verbal reports of the participants was carried out by two independent coders in two successive rounds. In the first round, both coders analyzed the participants' verbal reports independently. In the second round, each coder reviewed the interpretation of verbal reports for each follow-up probe from the other encoder. Finally, they held a meeting at which those verbal reports, in which there had been no agreement to whether or not they corroborated the previously established assumptions, were reviewed. The meeting ended with full agreement between the two coders.

3 Findings

The cognitive interview results have been divided according to the evidence obtained for each of the cases.

3.1 Assumption 1: Meaning of Family Concept

The analysis revealed that respondents had thought about different situations with family, relationships within the family, members of the same, about problems (accidents, labor problems, etc.) in which they had received help from family, etc. Example 1 shows some extracts of transcripts.

Example 1 Interpretation of “family support” construct

Interviewer: “Please tell me what you were thinking while responding to questions about emotional support”

5F40SM: “Mostly I thought about after my accident. Before the accident we had a good family relationship but later, it was much better. For me it’s like I’ve won the lottery with my family, they give me unconditional support …”

Interviewer: “Please tell me what you were thinking while responding to questions about emotional support”

11M65MM: “Well … no, I was thinking about my family situation … my oldest son is married and visits us every Saturday, the youngest is getting married this Friday … we are a united family … we try to stay in touch …”

Interviewer: “Please tell me what you were thinking while responding to questions about emotional support”

17M51MM: “… For my family it is a very important value, you must help all your brothers, your sisters, parents, grandparents … the family is important …”

Participant codes: (1) number of interview, (2) gender (*M* Male, *F* Female), (3) age, (4) marital status (*S* single, *M* married, *D* divorced) and (5) type of household (*S* single person and *M* multi-person)

In addition, participants maintained the same interpretation of the concept of “family” throughout the follow-up process. Respondents referred to the same people throughout the probes. The following transcript illustrates this evidence.

Example 2 Consistent meaning of family concept

Interviewer: “Could you tell me what were thinking while responding to questions about emotional support and your understanding of the role of the family?”

17M51MM: “Well of my family, my daughters and my wife.”

Interviewer: “How many people form your close family?”

17M51MM: “My wife and my daughters. Three.”

Interviewer: “And with whom in your family do you have regular contact?” (Probe for item 4: “Are you satisfied with the time you and your family spend together?”)

17M51MM: “With my wife and two daughters”

Interviewer: “And how much time do you and your family spend together?”(Probe for item 4: “Are you satisfied with the time you and your family spend together?”)

17M51MM: “My wife and my daughters… very little… 2 h a day, if possible, there are times when less. Sometimes nothing. There are times when I get home, my daughters have not arrived, I lie down and they come later. And sometimes maybe they arrive before, but I arrive later, and when I arrive are in bed. So… very little.”

Interviewer: “And which members of your family take those decisions together?” (Probe for item 3: “Are important decisions taken together?”)

17M51MM: “everyone, the four of us, the four. Because my daughters work…, my wife and I. The four”

Participant codes: (1) number of interview, (2) gender (*M* Male, *F* Female), (3) age, (4) marital status (*S* single, *M* married, *D* divorced) and (5) type of household (*S* single person and *M* multi-person)

These verbal reports corroborated the “Assumption 1”. They show that participants have agreed with the family support construct, they have understood the theme of the text of the items and keep a consistent interpretation of the concept of family throughout the items.

3.2 Assumption 2: Family Member Concept (People with Whom the Participants Live)

Analyses of verbal reports have shown that participants carry out an interpretation of family members different from that previously expected. That is, the participants included in their concept of family not only the people living together, but they also include other relatives and even friends. The following excerpts from transcripts are examples of this evidence.

Example 3 Family concept from people who live alone

Interviewer: “How many people are there in your immediate family?”

1M32SS: “Four, my parents and my two brothers”

Interviewer: “How many people are there in your immediate family?”

13F48SS: “There’s seven of us, and four nieces and nephews… eleven, and counting husbands and wives, then we’re fourteen”

Interviewer: “How many people are there in your immediate family?”

14M38SS: “four or five”

Interviewer: “And who are they?”

14M38SS: “My friends: Diego, Tomás, Jimena, Claudia and Fernanda”

Participant codes: (1) number of interview, (2) gender (*M* Male, *F* Female), (3) age, (4) marital status (*S* single, *M* married, *D* divorced) and (5) type of household (*S* single person and *M* multi-person)

3.3 Assumption 3: Ability to Grade ‘the the Family Support Construct’ to a Polytomous Answer System

The evidence obtained corroborates the ability of participants to grade their response and adjust it to multiple choice options. Example 4 illustrates the results with extracts of transcripts.

Example 4 Ability to grade answers

Interviewer: “In the question, ‘do you feel that your family loves you’, you replied ‘always’, in what situations would you answer ‘sometimes’?”

12M24SM: “Well, it would have to be something… like they’re ignoring me or not showing interest in me or there was a big break up. Something like that”

Interviewer: “In the question, ‘do you feel that your family loves you’, you replied ‘almost always’, in what situations would you answer ‘almost never’?”

16F54MM: “Well if I were to spend 3 days away from home without warning, and they didn’t call or worry, or anything; that would be a sign that you don’t mean anything to them and they don’t worry about you. And if they don’t worry about you, well… then they don’t love you, I suppose”

Participant codes: (1) number of interview, (2) gender (*M* Male, *F* Female), (3) age, (4) marital status (*S* single, *M* married, *D* divorced) and (5) type of household (*S* single person and *M* multi-person)

As can be seen, when the interviewer offers a choice of different responses participants are perfectly able to adjust their response to the hypothetical option that is offered. Thus, respondents are able to grade their responses in regard to the construct ‘family support’ and feelings received from their family, and adjust these responses to a polytomous response system.

4 Discussion

The aim of this study was to illustrate the use of cognitive interviewing method to gather validity evidence on response processes, based on the argument approach to validity. This objective has been achieved in the framework of the validation study of the APGAR scale on the family function construct. The following evidence of fit between the assumptions of the interpretive argument for the use of scores and response processes of the participants has been gathered.

With regard to ‘Assumption 1”, the participants answered the items of the construct based on “family support’ and keeping in mind a constant meaning for the “family concept” across APGAR items. The evidence on the response processes of the participants support the feasibility of the measuring instrument to access the construct “family support” and the unidimensionality proposed for the APGAR scale responses.

In relation to “Assumption 2”, respondents not only include the people living together in their concept of family, but also include other close relatives and even friends. This evidence shows the measurement errors that can occur when you insert the APGAR scale in a health survey as the sampling unit is households. Scores on the satisfaction of respondents with the support of their family include members from outside the household, other relatives or friends.

Third, in relation to “Assumption 3” verbal reports have shown that the construct “family support” of the participants can be graded and adjusted according to a polytomous response, confirming the usefulness of the multiple-choice design of the APGAR scale.

In general, the evidence found support the usefulness of the APGAR scale for measuring thee “family support” construct. However, the cognitive interviewing evidence does not support the intended use of the scale in the SNHS. The SNHS used the home as the sampling unit and try to limit the scope of the concept of family to be people with whom respondents live. Thus, by linking the APGAR scale data with the other variables of the SNHS, one is comparing different sample units and could mislead survey data users when drawing inferences from APGAR scale data.

Lastly, cognitive interviewing has offered useful information about the respondents’ response processes when responding to the items of a psychological scale, provided that that these response processes are not automatic, and that participants use information stored and available in the short term memory (Conrad and Blair 2009). Cognitive interview data have also revealed their ability to provide useful validity evidence of the response processes for the investigation of substantive aspects of construct validity (Messick 1995).

This study has sought to contribute to the discussion of methods for obtaining evidence of validity and how to interpret such evidence, specifically in the case of sources of evidence based on response processes (AERA et al. 1999). Using the argument-based approach to validity that Michael T. Kane proposed in 1992 and 2006, when specifying, we focused on the pattern of inferences and assumptions on which the users of the scale would rely on to interpret scores and implement decision-making. From there, an “interpretive argument” was created and the cognitive interview method applied to examine the assumptions that support the desired interpretation of the scores (Kane 1992). This form of approach to the validity studies clarifies which evidence sources to extract information from and how to conduct the validation study, with the consequent advantage of eliminating the feeling of frustration and loss to which Sijtsma (2009) and Zumbo (2009) refer, respectively.

On the other hand, there are some doubts about the role of construct validity evidence in the Kane's argument-based approach to validity. For example, Chapelle et al. (2010) state that one of the strengths of Kane's approach is that there is no need for construct validity evidence. From our perspective, the point is whether or not the interpretative argument behind the intended use of test score relies on assumptions about the processes involved in task measurements. Assumptions about how people interpret and respond to items and scale require construct validity evidence. This view is compatible with more comprehensive approaches to validity like Zumbo's model of contextualized explanation (Zumbo 2009). Provided that the score interpretations by users of tests scales rely on assumptions about respondents' response processes, score explanations can be elaborated when conducting a validation study following Kane's approach to validity. Besides, cognitive interviewing can bring together both approaches to validity providing validity evidence for the context of the explanation and the context of assessments from "experiential experts" (Gadermann et al. 2011), which carry their educational and social backgrounds to the validation processes.

Moreover, from a methodological point of view this study has illustrated how the cognitive interview method designed and used for evaluation of the questionnaires inserted in surveys can also be used in the evaluation of psychological scales included in surveys. Although it is easy to assume that participants develop similar response processes while answering questions from questionnaires and questions of psychological scales, the pre-test methods are very poorly applied to the evaluation of these scales. We believe that this is due to the mistaken belief that the metric properties are inherent in the psychological scales, taking the validity of their measurements for granted, assuming that these metrical properties will not vary once the measuring context has completely changed, and thus, the application of such scales. From this perspective, survey researchers can make valuable contributions to change this established misconception (Willis 2005).

Acknowledgments This study was funded by the Andalusia Regional Government under the Excellent Research Fund (Project n° SEJ-5188).

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Beatty, P., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly.* doi:10.1093/poq/nfm006.

Bellón, J. A., Delgado, J., Luna, P., & Lardelli, P. (1996). Validez y fiabilidad del cuestionario de función familiar Apgar-familiar. *Atención Primaria, 18*, 289–296.

Bickart, B., & Felcher, M. (1996). Expanding and enhancing the use of verbal protocols in survey research. In N. Schwarz & S. Sudman (Eds.), *Answering question. Methodology for determining cognitive and communicative processes in survey research* (pp. 115–142). San Francisco: Jossey-Bass.

Boeije, H., & Willis, G. B. (2011). *The cognitive interviewing reporting framework (CIRF): Incorporating the principles of qualitative research.* Paper presented at the Fourth Conference of the European Survey Research Association (ESRA), Lausanne, Switzerland. Abstract retrieved from http://surveymethodology.eu/conferences/lausanne-2011/presentation/226/.

Castillo, M., Padilla, J. L., Gómez, J., & Andrés, A. (2010). A productivity map of cognitive pretest methods for improving survey questions. *Psicothema, 22*(3), 475–481.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice.* doi: 10.1111/j.1745-3992.2009.00165.x.

Conrad, F. G., & Blair, J. (2009). Sources of error in cognitive interviews. *Public Opinion Quarterly.* doi: 10.1093/poq/nfp01.

Conrad, F. G, Brown, N. R., & Cashman, E. (1998). Strategies for estimating behavioural frequency in survey interviews. *Memory.* doi:10.1080/741942603.

DeVellis, R. F. (2003). *Scale development: Theory and application* (2nd ed.). Thousand Oaks, CA: Sage.

DeWalt, D. A., Rothrock, N., Yount, S., & Stone, A. A. (2007). Evaluation of item candidates the PROMIS qualitative item review. *Medical Care.* doi:10.1097/01.mlr.0000254567.79743.e2.

Ding, L., Reay, N. W., Lee, A., & Bao, L. (2009). Are we asking the right questions? Validating clicker question sequences by student interviews. *American Journal of Physics.* doi: 10.1119/1.3116093.

Ercikan, K., Arim, R., & Law, D. (2010). Application on think aloud protocols for examining and confirming sources of differential item functioning identified by experts review. *Educational Measurement: Issues and Practices, 29*(2), 24–35.

Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2011). Investigating the substantive aspect of construct validity for the satisfaction with life scale adapted for children: A focus on cognitive processes. *Social Indicator Research*. doi:10.1007/s11205-010-9603-x.

Huber, G. L. (2008). *AQUAD, software for the analysis of qualitative data (version 6.8.1.1)*. Ingeborg Huber Verlag, University of Tübingen, Germany.

Irwin, D. E., Varni, J. W., Yeatts, K., & DeWalt D. A. (2009). Cognitive interviewing methodology in the development of a pediatric item bank: a patient reported outcomes measurement information system (PROMIS) study. *Health and Quality of Life Outcomes.* doi:10.1186/1477-7525-7-3.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*. doi:10.1037/0033-2909.112.3.527.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Washington, DC: American Council on Education/Praeger.

Kane, M. (2010). Validity and fairness. *Language Testing.* doi: 10.1177/0265532209349467.

Knaff, K., Deatrick, J., Gallo, A., Holcombe, G., Bakitas, M., Dixon, J., & Grey, M. (2007). The analysis and interpretation of cognitive interviews for instrument development. *Research in Nursing and Health.* doi:10.1002/nur.20195.

Lissitz, R. (2009). *The concept of validity: revisions, new directions and applications*. Charlotte, NC: Information Age Publishing In.

Llosa, L. (2008). Building and supporting a validity argument for a standards-based classroom assessment of English proficiency based on teacher judgments. *Educational Measurement: Issues and Practice.* doi: 10.1111/j.1745-3992.2008.00126.x.

Menon, G. (1993). The effects of accessibility of information in memory on judgments of behavioral frequencies. *Journal of Consumer Research, 20*(3), 431–440.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 3–103). Washington, DC: American Council on Education.

Messick, S. (1995). Validity of psychological assessments: Validation of inferences from person's responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749.

Olt, H., Jirwe, M., Gustavsson, P., & Emami, A. (2010). Psychometric evaluation of the swedish adaptation of the Inventory for Assessing the Process of Cultural Competence among Healthcare Professional Revised (IAPCC-R). *Journal of Transcultural Nursing.* doi:10.1177/1043659609349064.

Padilla, J. L., Gómez, J., Hidalgo, M. D., & Muñiz, J. (2007). Esquema conceptual y procedimientos para analizar la validez de las consecuencias del uso de los test. *Psicothema, 19*(1), 173–178.

Poole, H. M., Murphy, P., & Nurmikko, T. J. (2009). Development and preliminary validation of the NePIQoL: A quality-of-life measure for neuropathic pain. *Journal of Pain and Symptom Management.* doi:10.1016/j.jpainsymman.2008.01.012.

Sijtsma, K. (2009). Correcting fallacies in validity, reliability and clasification. *International Journal of Testing*. doi:10.1080/15305050903106883.

Sireci, S. G. (2005). Validity theory and application. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 2103–2107). Nueva, Jersey: Wiley.

Sireci, S. G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 19–37). Charlotte, NC: Information Age Publishing Inc.

Smilkstein, G. (1978). The family APGAR: A proposal for family function test and its use by physicians. *Journal of Family Practice, 6*, 1231–1239.

Smilkstein, G., Ashworth, C., & Montano, D. (1982). Validity and reliability of the family APGAR as a test of family function. *Journal of Family Practice, 15*, 303–311.

Spanish Ministry of Health and Consume. (2006). Encuesta Nacional de Salud de España 2006 [National Health Survey in Spain, 2006]. Madrid: Ministerio de Sanidad y Consumo. http://www.ine.es/jaxi/menu.do?L=0&type=pcaxis&path=%2Ft15/p419&file=inebase (March, 13 2008).

Tourangeau, R. (1984). Cognitive science and survey methods: A cognitive perspective. In T. Jabine, M. Straf, J. Tanur, & R. Tourangeau (Eds.), *Cognitive aspects of survey design: Building a bridge between disciplines* (pp. 73–100). Washington, DC: National Academy Press.

Tourangeau, R., Rips, R. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.

Willis, G. B. (2005). *Cognitive interviewing*. Thousand Oaks: Sage Publications.

Willis, G. B., Royston, P., & Bercini, D. (1991). The use of verbal report methods in the development and testing of survey questionnaires. *Applied Cognitive Psychology, 5*, 251–267.

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 45–79)., Psychometrics Amsterdam: Elsevier.

Zumbo, B. D. (2009). Validity as contextualized and pragmatic explanation, and its implications for validation practice. In R. W. Lissitz (Ed.), *The concept of validity: revisions, new directions and applications* (pp. 65–82). Charlotte, NC: IAP—Information Age Publishing Inc.

Zumbo, B. D., & Shear B. R. (2011). *The concept of validity and some novel validation methods*. Paper presented at the 42nd Annual Meeting of the Northeastern Educational Research Association, Rocky Hill, CT. Abstract retrieved from http://www.nera-education.org/Full_2011_NERA_Program.pdf.