Psicothema

Psicothema 2014, Vol. 26, No. 1, 136-144
doi: 10.7334/psicothema2013.259

ISSN 0214 - 9915 CODEN PSOTEG

www.psicothema.com

Validity evidence based on response processes

José-Luis Padilla and Isabel Benítez
University of Granada

Abstract

Background: Validity evidence based on response processes was first introduced explicitly as a source of validity evidence in the latest edition of Standards for Educational and Psychological Testing. In this paper, we present the theory, the relationship with other sources of validity evidence, and the methods available for validation studies aimed at obtaining validity evidence about response processes. **Method:** A comprehensive review of the literature along with theoretical and practical proposals. **Results:** The articles provides arguments for determining when validity evidence based on response processes is critical for supporting the use of the test for a particular purpose, and examples of how to perform a validation study to obtain such validity evidence. **Conclusions:** There are methods for obtaining validity evidence based on response processes. Special attention should be paid to validation studies using the cognitive interview method given its features and possibilities. Future research problems pose how to combine data from different methods —qualitative and quantitative—, to develop complete validity arguments that support the use of the test for a particular purpose.

Keywords: Validity, standards, evidence of response processes, cognitive interviewing.

Resumen

Evidencia de validez basada en los procesos de respuesta. **Antecedentes:** la evidencia de validez basadas en los procesos de respuestas fue incluida explícitamente por primera vez como fuente de evidencias de validez en la última edición de los *Standards for Educational and Psychological Testing*. En este artículo, presentamos la teoría, la relación con otras fuentes de evidencias de validez, y los métodos disponibles para realizar estudios de validación cuyo objetivo sea obtener evidencias de validez sobre los procesos de respuesta. **Método:** una extensa revisión de la literatura junto con propuestas teóricas y prácticas. **Resultados:** el artículo aporta argumentos para determinar cuando la evidencia de validez basada en los procesos de respuesta es crítica para apoyar el uso del test para un objetivo particular, y ejemplos de cómo realizar un estudio de validación para obtener tales evidencias de validez. **Conclusiones:** hay métodos para obtener evidencias de validez basadas en los procesos de respuesta. Debe prestarse especial atención a los estudios de validación mediante el método de entrevista cognitiva por sus características y posibilidades. Futuros problemas de investigación plantean como combinar datos de métodos diferentes —cualitativos y cuantitativos—, para elaborar argumentos de validez que apoyen el uso del test para un objetivo particular.

Palabras clave: validez, standards, evidencias de procesos de respuesta, entrevista cognitiva.

The study of the response processes to test items and questionnaires was first considered explicitly as a source of validity evidence in the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999). The earlier version of these *Standards* (APA, AERA, & NCME, 1985) included the study of "individual response processes" as part of evidence related to construct validity. However, the previous editions in 1954, 1966, and 1974 did not consider response processes as "types," "aspects," or "categories" of validity. In fact, there are no detailed references to the study of response processes as evidence of validity in the classic articles of validity theory, widely cited in the literature (e.g., Kane, 1992; 2006); with the exception of the seminal works of S. Messick anticipating the role of the validity evidence based on response processes (e.g., Messick, 1989; 1990). Messick (1990) included among the "forms of validity evidence": "Directly probe the ways in which individuals cope with the items or tasks, in an effort to illuminate the processes underlying item response and task performance" (p. 5). In general terms, the absence or oversight justifies the benchmark role that the 1999 *Standards* played in determining the content, methods and scope of the source of validity based on response processes.

Prior to the appearance of the 1999 *Standards*, lines of research in the field of educational testing had been developing, such as those of Embretson (1983) connecting item response theory (IRT) and cognitive psychology and Mislevy, Steinberg and Almond (2002) on model-based reasoning for the development of tasks and items, which could be included within the field of validation based on response processes. In fact, it could be said that interest in explaining the processes of responses to test items has been around since the origin of validity. Sireci (2009) criticized earlier definitions of validity as the degree to which "... a test measures

Received: August 27, 2013 • Accepted: September 27, 2013
Corresponding author: José-Luis Padilla
Faculty of Psychology
University of Granada
18071 Granada (Spain)
e-mail: jpadilla@ugr.es

what it is supposed to measure" (Garrett, 1937, p. 324). The key is in the clause "what it is supposed to measure." He argued it is not possible to properly interpret the score on a test, if you do not know what the test measures. "Knowing what the test measures" is a critical aim of many validation studies; and throughout the evolution of validity theory, the vision of validity by those who advocate a substantive theory of response processes (e.g., Borsboom, Mellenbergh, & van Heerden, 2004), or those who understand validity as a contextualized and pragmatic explanation of the scores on the test (e.g., Zumbo, 2009) can be traced.

In this paper, we (1) describe validity evidence based on response processes and its relation to other sources of validity evidence, (2) determine when it is critical to have evidence based on the response process to support the use of the test for a particular purpose, and (3) present methods available for conducting validation studies on response processes, with special attention to the cognitive interview method. Throughout the article, we will also present studies to illustrate the contents and trace the practice of validation studies focused on this source of validity evidence.

Validity evidence of the response processes in the standards 1999

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). The *Standards* provide few indications for obtaining evidence about the response processes:

> Questioning test takers about their performance strategies or response to particular items… Maintaining records that monitor the development of a response to a writing task… Documentation of other aspects of performance, like eye movement or response times (p. 12).

The indications for obtaining evidence about response processes mix who provides the data ("test takers"), with the data itself ("records"). This confusion has made it difficult to develop methods aimed at obtaining this source of evidence. In addition, the *Standards* state that the study of response processes can go beyond the test takers or examinees to include "... observers or judges to record and/or evaluate examinees' performance or products" (p. 13). This call opened the study of response processes for judges or observers in areas of educational and psychological testing such as standard-setting, subject-matter experts for content validity, scoring of performance assessments, etc. Such studies are especially important when the validity argument that justifies using the test requires evidence that the appropriate criteria are being used, and the observers or judges are not being influenced by factors irrelevant to the intended interpretation.

Relationship between evidence based on test content and response processes

The *Standards* emphasize that the sources of evidence do not represent different types of validity, since "Validity is a unitary concept" (p. 11), but rather they illuminate different aspects of validity. One of the ideas presented in this paper is to strengthen the complementarity between sources, especially in the case of sources of evidence based on test content and those based on response processes. Complementarity in the objectives and, as discussed below, in the areas of validity they face.

Validity based on test content comes from the analysis of the relationship between the content of the test and the construct meant to be measured. The key to understanding the "complementarity" in both sources is in what the *Standards* describes as "test content": "Test content refers to the themes, wording, and format of the items, tasks, or questions on a test" (p. 11). Sireci & Faulkner-Bond (2014) in this issue summarizes how evidence based on test content comes from logical or empirical analysis of the adequacy with which the test content represents the content domain and its relevance to the proposed interpretation of the scores in the test, almost exclusively through the use of "subject-matter experts." The question is how to judge the relevance without resorting to evidence based on response processes to provide data about "…the detailed nature of the performance actually engaged in by the test takers or examinees" (p. 12) as *Standards* state. The validation process presented by Sireci (2012) to support the use of the Massachusetts Adult Proficiency Tests (MAPT), clearly illustrates, as discussed below, the complementarity of the two sources.

In support of complementarity between sources, it is curious how the *Standards* repeat indications on how to use both sources of evidence to address "... questions about difference in meaning or interpretation of the test scores across relevant groups of examinees" (p. 12). The indication points to the combined use of both sources to address the detection and interpretation of differential item functioning.

Standards and testing issues for evidence based on response processes

The relationship between the sources of validity evidence based on test content and response processes are more apparent when identifying standards and testing issues for which they are relevant.

Standard 1.8 indicates that:

> If the rationale for a test use or score interpretation depends on premise about the psychological processes or cognitive operations used by examinees, then theoretical or empirical evidence in support of those premises should be provided (p. 19).

After extending the application to observers, the commentary accompanying standard 1.8 supports the thesis of complementarity between the evidence based on content and response processes. This comment requires the provision of evidence about the psychological processes or cognitive operations when the test specifications determine the process to be evaluated.

Under the heading "Bias associated with test content and response processes" the AERA et al. (1999) *Standards* point to the two sources of evidence to look for causes of bias in the tests. The explanation of the bias is reflected in the concept of "construct-irrelevant variance" appropriate in the case of test content when there has been an inadequate sampling of content, and in the case of the response process when "... test items elicit varieties of response other than those intended or can be solved in ways that were not intended "(p. 78). The examples listed below are "classic" examples in the presentation of test bias and differential item functioning: differential trend acquiescence, different familiarity

with item response formats, and where the performance on items depends on some ancillary ability which the comparison groups have to different degrees.

Considering the role that *Standards* point out for evidence based on response processes in the treatment of bias in tests, it should not be surprising that they include references to said source when addressing the issue of "testing individuals of diverse linguistic backgrounds" (e.g. standard 9.2), or "testing individuals with disabilities" (e.g. standard 10.1).

When should a validation study based on response processes be conducted?

It is beyond the scope of this article to summarize the exciting debate on validity theory, a discussion that emerged in the commencement of evaluation through tests and questionnaires and which is still very active today. However, proposing when evidence based on response processes is critical requires at least a lay out of the perspectives and main contents about validity.

A reading of the most recent theoretical work about validity indicates a majority and basic consensus on the current contents of the Theory of Validity, even among those who defend different views (e.g., Cizek, 2012; Kane, 2006, 2013; Sireci, 2009, 2012; Zumbo, 2009). Put simply, the arguments of consensus about validity are: (1) it belongs to the "entitled" inferences and interpretations for the use of the test, (2) it is not a characteristic of the test or questionnaire, (3) it is a unitary concept, and (4) it is an evaluative judgment. The idea of differentiating between the concepts of "validity" and "validation" is also shared, the latter referring to the methods and, in particular, the process for obtaining evidence with which to support test use. However, there are some differences between those who direct the validation towards the "particular use of test" (Sireci, 2009, 2012), "the assumptions that underpin the interpretive argument" (Kane, 2013), or the "explanation" of the differences between the scores on the test (e.g., Zumbo, 2009).

Given the practical focus of the present article and the reference to the AERA et al. (1999) *Standards* as a framework for validation studies, the answer to the question of when evidence based on response processes is critical fits into what Sireci (2012) calls a "de-constructed approach to test validation" aimed at providing the necessary validity evidence to support the use of the test. The answer is also partly addressed by the "argument-based approach to validation" made by Kane (2006).

In both cases, the answer to the question is supported by two pillars: first, the concept formulated by Embretson (1983) of "construct representation" which includes as threats to validity those from a "construct under representation," and those from "construct-irrelevant variance;" and secondly, the development of "rival hypothesis" that as indicated by the AERA et al. (1999) *Standards* defy the proposed interpretation to justify the use of the test.

On these two conceptual pillars, practitioners should assess the performance of a validation study to obtain evidence based on response processes that justify the particular use of a test when, among the propositions that have been made to justify the use of the test:

a) The performance of the "test takers" or "examinees" in the test or questionnaire items reflects the psychological processes and / or cognitive operations delineated in the test specifications.
b) The processes of judges or observers when evaluating the performance or products of the different test takers are consistent with the intended interpretation of the scores.
c) Groups of test takers defined by demographics, linguistic or other conditions associated with the intended use of the test, did not differ in the nature of their performance or in the responses because of sources of "construct-irrelevant variance."

Addressing the validation of response processes of judges and observers has been explored for instance in the standard-setting area. In general, the studies done aimed to reveal the cognitive process of panelist (e.g., Skorupski & Hambleton, 2005), or the factors that influenced panelist's decisions (e.g., Ferdous & Plake, 2005). Two examples may illustrate how to address the validation of the propositions a) and c).

Sireci (2012) proposed a scheme of validation developing several studies to justify the use of the Massachusetts Adult Proficiency Tests (MAPT). The whole process of development and validation of MAPT is in the MAPT technical manual (Sireci et al., 2008). The MAPT was developed for a statewide assessment focused on adult basic education. The particular focus of MAPT is "... to measure ABE learners' knowledge and skills in math and reading so that their progress in meeting educational goals can be evaluated" (p. 8). The goal statement led to the identification of six general validity questions that guided the validation studies. Of these, there is a general validity question whose answer implied conducting a validation study aimed at obtaining evidence about response processes: "Do MAPT scores provide accurate information regarding [adult basic education] students' math and reading proficiencies?" "Accurate" should be understood in terms of proposition a) stated above: Do the examinees respond to the MAPT according to the cognitive operations delineated in the specifications of the test? Sireci et al. (2008) addressed the validation study from "rival hypothesis": Are examinees guessing without reading the items, are they seriously engaged in responding to the test, or do they have enough time to respond. The data for examining the hypotheses were measurements of response time. Overall, the results allowed the researchers to reject the rival hypotheses.

Castillo and Padilla (2012) conducted a validation study using cognitive interviews following the argument based approach to validation (Kane, 2006), to obtain evidence of the processes of response to the items of a psychological scale designed to measure the construct "family support." The validation study sought to examine the assumption that people living alone and those living with others, responded to the items using the same psychological processes. The comparison between the psychological processes of both groups did not support this assumption.

Practice of validation studies for validity evidence of the response processes

The analysis of the source of validity evidence based on response processes leads to reviewing the practice of such studies: how frequently they are carried out, what their goals are and what evaluation contexts have been conducted since the AERA et al. (1999) *Standards* included this source of evidence. As we shall

see, the picture is not very encouraging. Compared to the other sources, validation studies aimed at obtaining evidence from response processes are scant. The resistance of the researchers to modifying their vision and practices in the validation studies, the novelty with respect to the inclusion of this source of validity, together with the absence of clearer advice on how to obtain evidence from response processes, may explain the situation.

Some studies have attempted to review the practice of validation studies. For example, Cizek, Rosenberg, and Koons, (2007) reviewed validity papers to evaluate whether the validation framework used by authors fit what they called "modern validity theory" and what the habitual trends were in conducting validation studies. They found the majority of the papers were far from the modern and unitary concept of validity. Validity was, in a many cases, not clearly defined and also considered a test characteristic. In relation to the sources of validity evidence investigated, the authors found the majority of the papers were focused on content and construct validity, and validity evidence based on participants' response processes were studied only in 1.8% of the papers. Sireci and Parker (2006) compared the conceptualization of validity as proposed in *Standards* with validity evidence presented in in the courtroom. They concluded testing agencies were close to the *Standards* indications although none of the studies reviewed put forward evidence based on response processes. Also, Zumbo and Shear (2011) showed a higher presence in the medical outcomes field, where 14% of the validation studies were based on participants' response processes. In other fields, very few studies were focused on evidence based on respondents' response processes; however the number increased between 2000 and 2010.

With the aim of contributing and somehow updating the description of the practice in the validation studies based on response processes, we performed a literature search focused on the journals indexed in the ISI Web of Knowledge, starting from the publication of AERA et al. (1999) *Standards*. The main objectives of the search were to characterize available validation studies in terms of the assessment context in which they were made, their objective and the method used. Keywords that guided the search were a pool of terms that refer to the source of validity based on response processes ("validity evidence based on response processes," "cognitive processes," "validity"), plus terms with which the *Standards* refer to how to obtain this evidence is obtained ("interview," "eye tracking," "response times"). As expected, the combined use of keywords was more effective in finding articles in which validation studies on response processes were presented.

Only 63 papers were selected after reading the abstract to confirm that the object and the content of these responded to validation studies on response processes. The small number of papers is in line with those found in the review articles cited above. The first contribution of the search to the characterization of the practice of validation studies is due to some of the difficulties encountered. For example, the terms "sources of evidence based on response processes" or "evidence of response processes," does not identify all the validation studies that aim to investigate the psychological processes or cognitive operations of test takers when responding to the items. Neither was it very useful to identify the articles based on the method used. As the indications of the AERA et al. (1999) *Standards* are not too explicit, not even combining the method name with labels referring to the source of validity guaranteed to identify articles focused on this source of evidence.

Regarding the specific findings, the review of selected articles revealed that a large majority are validation studies in the area of health during the development of scales for specific diseases or health processes. These results coincide with that of Zumbo and Shear (2011). For example, Deal, DiBenedetti, Williams, and Fehnel (2010) elaborated and validated a pain scale; Althof, Perelman, and Rosen (2011) followed the same process with a sexual arousal scale; and Brod, Hammer, Christensen, Lessard, and Bushnell (2009) implemented interviews for discovering patients perspectives about diabetes. However, some papers were also found that focused on "psychological processes" such as processing speed (Cepeda, Blackwell, & Munakata, 2013), or decision strategies (Day, 2010).

It is also significant that the validation studies on response processes are often carried out during the development of the test or questionnaire. For example, Gehlbach and Brinkworth (2011) proposed a framework for developing scales in which cognitive interviews and focus groups are proposed as procedures for assuring the quality of the instrument elaborated. To illustrate their proposal, authors used scales to assess teacher–student relationships (TSR) from both teachers' and students' perspectives at the middle- and high-school levels. Deal et al. (2010) also searched for validity evidence during the elaboration process but following an interactive process. These authors created a questionnaire for assessing Endometriosis pain and bleeding diary by conducting an expert panel and focus groups, later the first draft was evaluated by cognitive interviews which were conducted in three rounds. Evidence obtained was included in the questionnaire and later a pilot study was conducted for finally assessing psychometric properties.

Validation studies based on response processes during the evaluation of the psychometric properties of the test and/or questionnaire are also illustrative. Most of these studies seek to identify elements of the items that can cause mismatches between responders' psychological processes and those delineated in test specifications. For example, Olt, Jirwe, Gustavsson, and Emami (2010) evaluated the psychometric properties of a questionnaire intended to assess the cultural competence among healthcare professionals. They related difficulties reported by healthcare professionals in understanding the construct of "cultural competence" with poor evidence of questionnaire reliability and internal structure.

Methods for obtaining validity evidence of the response processes

Describing the status of the validation studies focused on evidence of response processes requires a review of the methods used. The presentation of the methods is particularly relevant in the analysis of this source of validity evidence, since the *Standards* themselves, as already noted, indicate from where such evidence may be derived, at least in the case of the analysis of responses to items in tests and questionnaires from individual examinees: Questioning test takers, records that monitor the development of a response, eye movement, and response times. From these indications, the methods have been grouped into two categories: those that directly access the psychological processes or cognitive operations (think aloud, focus group, and interviews), compared to those which provide indirect indicators which in turn require additional inference (eye tracking and response times). Beginning

with the presentation of each method for the latter, after which so-called direct methods are presented, and finally, a more detailed description of the cognitive interview method.

Response times

Validation studies that measure response times are habitually focused on connecting response time with the complexity of processes involved in developing the task (Cepeda et al., 2013). Response times validation studies seek to obtain evidence of the response processes or some aspect of them (e.g., guessing, commitment to responding to items, etc.), by registering response times while test takers are responding to the items. The validation study done by Sireci et al. (2008) for the MAPT is a good example as seen in previous sections. Wang and Sireci (2013) also found a relationship between the complexity of the cognitive operations involved with the items and the time that examinees took to respond to them, and how the relationship was provoked and mediated by item difficulties.

Eye-tracking methods

Eye-tracking or eye-movement has also been used as indirect cues to attention and cognitive process (Day, 2010). Usually, eye-tracking methods are implemented during the task of responding to test or scale items to gain access to the psychological processes or cognitive operation involved. For example, Ivie and Embretson (2010) applied the eye-tracking method for obtaining evidence about cognitive processes involved in the assembling object items. Elling, Lentz, and de Jong (2012) resorted to eye tracking to validate concurrent think-aloud protocol.

Interviews

As indicated earlier, in general the interview is researchers' preferred method in validation studies based on response processes. The preference is understandable as this method responds directly to the AERA et al. (1999) *Standards'* guidelines, asking respondents to items about their psychological processes and cognitive operations. It is also a method that is easy to apply and requires relatively few resources. Furthermore, the literature reflects a variety of names to refer to similar procedures in-depth interviews, semi-structured interviews, think-aloud protocols, etc. All of these were grouped under the label "interview," and the term "cognitive interview" was left to describe the type of interview advocated in this article and described in detail later.

The interview method acquires different nuances depending on whether the goal is to identify elements (words, expressions, response format, etc.), which may be problematic for the test or questionnaire respondents, or if the researcher intends to identify how people refer to the object, content, or specific aspects included in the items. In both cases, the aim is that the items do not hinder the fit between the response processes and those delineated in the test specifications. For example, Krall and Lohse (2010) were interested in the eating competence of women in a program for nutrition and education assistance because they had evidence this competence can be especially affected in specific situations. They designed interviews to validate the ecSatter Inventory for low-income women. Brod et al. (2009) conducted interviews to assure participants were interpreting concepts around diabetes as defined in the test specifications of the Treatment Related Impact Measure-Diabetes. Information from participants was analyzed by classifying emerging themes and concepts, which allowed researchers to propose and implement changes in the measures being administered. Althof et al. (2011) applied interviews to identify useful elements for elaborating the Subjective Sexual Arousal Scale for Men, and later again applied interviews in a second phase of the study to check that the participants' interpretation, thought processes and/or feelings fitted to the intended construct.

Think-aloud protocols and vignettes are sometimes used within or as key components of the interview method in validation studies. For example, Gadermann, Ghun and Zumbo (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, Arin and Law (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.

When using vignettes, researchers present short stories to the participants who should make a judgment about the situation illustrated (e.g., Martin, 2004). Participants should also indicate how and why they would respond in the situation described. Vignettes are especially useful for evaluating whether participants respond to the complete item or focus only on specific elements while responding to items.

Focus groups

Other methods are available that help gather evidence of response processes directly from participants, for example, focus groups. The focus group is considered a useful method for exploring unknown topics through group discussion about the topic, element, aspects, etc., included in the test or scale items (Hawthorne et al., 2006). Participants in the focus group can discuss feeling, thoughts, opinions, etc., due to the facilitator effect of the social interaction that can reveal their psychological processes while responding to the items. For example, Webber and Huxley (2007) implemented focus groups to assure the relevance of items and perspectives of participants in a scale for evaluating social capital.

Cognitive interviewing for obtaining validity evidence based on response processes

The cognitive interview method may be especially useful for gathering validity evidence based on response processes. The aim of this section is to present the logic of the method, point out its conceptual foundations, and present the most relevant features practitioners should consider when applying the method in a validation study. A detailed presentation of all the particulars involved in the application of the cognitive interview method is beyond the scope of this article. Manuals are available in the literature with different approaches but which detail the method step by step (e.g., Willis, 2005; Miller, Cheep, Wilson, & Padilla, 2013).

The aim of cognitive interviewing is to access the participants' cognitive processes, which are also gathered to provide validity evidence based on participants' response processes. Cognitive Interviewing (CI) is the most used cognitive pretest method in survey research when survey question developers and evaluators seek to understand the "question-and-answer" cognitive process carried out by the respondents when answering survey questions (Castillo, Padilla, Gómez-Benito, & Andrés, 2010). Beatty and Willis (2007) who described cognitive interviews as "the administration of draft survey questions while collecting additional verbal information about the survey responses, which is used to evaluate the quality of the response or to help determine whether the question is generating the information that its author intends" (p. 288).

The reason we advocate using CI for validation studies when evidence based on response processes is needed, is that this method can provide evidence about the extent to which psychological processes and cognitive operations performed by the respondents actually match those delineated in the test specifications.

Theoretical foundations of cognitive interviews

The growing use of the cognitive interview method in studies validating tests and questionnaires continues a long history of "relations" between the survey methodology ("survey research"), and psychometrics. Since its appearance, researchers from both disciplines recognize that the quality of the measurements determines the quality of social and psychological research. Throughout history there has been "borrowing" of concepts, methods and a common vision of errors grouped as random and systematic error. Both methodological contexts have addressed until recently the measurement error from a purely statistical model, and have also shared the main limitation identified by Tourangeau, Rips, and Rasinski (2004) in the case of survey research: focusing on the consequences rather than the causes of errors, without having generated significant knowledge about their origin and how to prevent them.

The development of the cognitive interview method is historically tied to what in research survey movement is known by the name of Cognitive Aspects of Survey Methodology (CASM). This movement emerged as a result of two conferences: the Advanced Research Seminar on Cognitive Aspects of Survey Methodology held in the United States in 1983 and the Conference on Social Information Processing and Survey Methodology which took place in Germany in 1984 (Jabine, Straf, Tanur, & Tourangeau, 1984). Both conferences stressed the importance of taking people into account as "active agents" and considering the cognitive processes involved in the response process. Tourangeau (1984) further developed the "question-and-answer" process model that included a description of the cognitive processes involved in the "question-and-answer" process. Figure 1 shows a model in which the cognitive processes that appear between the formulation of the question and the statement of the response are positioned.

Figure 1 shows the four phases that a person would complete when responding to a question. "Translating" the model into the context of testing, test takers go through four phases while responding to items developing the following cognitive operations: first, they interpret and understand the item or task ahead which involves understanding both the intended purpose and the concepts and expressions that are included, and then they retrieve the information needed to answer the question, then make a judgment that allows them to integrate and evaluate the information retrieved, and finally they adjust their response to the proposed alternatives and communicate it.

In the last two decades, the model has evolved to include the possible non-sequentiality of the phases in all circumstances (Collins, 2003), and the presence of motivational, social and cultural dimensions (Krosnick, 1999). These extensions of the models could be very useful for obtaining evidence of the response processes of not only those about cognitive operations, but also examinees' expectations, attributions, their life experiences what they bring when responding to test items.

Conducting cognitive interviews

The cognitive interview method can be described starting with the major steps and decisions the researcher must take when performing the validation study. Below, we summarize the contents of each step, with reference to their application for obtaining evidence based on response processes.

Sampling

The first decisions to be taken when applying the method refer to how many and who should be interviewed. It is not easy to understand the answer, if it is forgotten that the CI is a qualitative method whereby the "representativeness" of the participants is not a primarily numerical matter. The usual practice is to select the participants thinking about the characteristics that are relevant to the aim of the study. The AERA et al. (1999) *Standards* reiterate the idea of comparing response processes using "relevant subgroups of examinees." The definition of subgroups depends on the content of the proposition that the validation study intends to test. For example, Castillo and Padilla (2012) defined the groups according to whether people lived alone or with others. One would proceed similarly if the proposition being validated involves comparing subgroups defined by demographics, language, culture, etc.

In addition to previous considerations, in relation to the total number of participants in cognitive interviews, two criteria that come from the qualitative character of the CI should be taken into account: theoretical saturation and relevance. Both criteria affect the sampling because the number and characteristics of interviewees depend on the analysis of the interviews. Theoretical saturation means that researchers should keep conducting

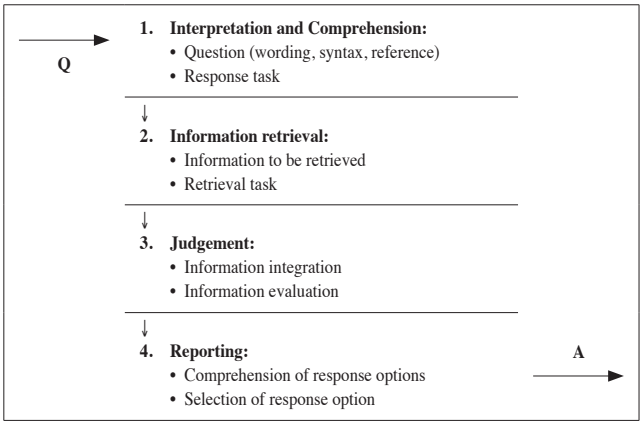

Figure 1. *Representation of the question-and-answer model*

interviews until no new "findings" emerge from the interviews, while "theoretical relevance" refers to selecting interviewees based on their theoretical relevance to the emerging findings. Usually, both criteria are met with a number between 20 and 50 interviews in terms of the objectives of the validation study and the complexity of the proposition to be tested.

Developing the cognitive interviewing protocol

The CI is a semi-structured interview in which the interviewer uses an interview protocol. The protocol guides the interview and, therefore, has a key role in implementing the method. Via the protocol, the interviewer asks the relevant questions to access the psychological processes and cognitive operations of the "test takers," getting the interviewee to assume the "role" required by the method, and at the same time, having the flexibility to obtain all relevant data on response processes.

The technical term for questions in the cognitive interviewing protocol is "probes" or "follow-up probes." In general terms, two main strategies can be identified for developing probes when conducting CI in educational and psychological testing: a think-aloud method focused on the verbalization of the thoughts of participants as they respond to test or scale items; or the probe based method, which develops follow-up probes for specific areas of each item. In the latter case, which is closest to the approach proposed in this paper, the probes are developed based on the features or elements of the items that researchers consider potentially problematic due to examinees or relevant subgroups of examinees interpreting the items differently. Different types of follow-up probes were proposed by Willis (2005) for covering different objectives. These probes, which also provide different types of evidence, are illustrated in Table 1. All the probes in the example were designed to investigate the response processes to the question "How was your health in the last twelve months?"

The second key element of the CI protocol is instructions by means of which the interviewer presents the study to the interviewees, with the aim of explaining what their role in the study is supposed to be. Figure 2 shows an example of part of the instructions provided to the interviewees by Benitez and Padilla (2013). The aim of the study was to obtain validity evidence about the response processes associated with DIF in a cross-lingual study. The author compared the response processes of the US and Spanish students to the items of the PISA 2006 Student Questionnaire.

Table 1
Types and examples of follow-up probes

Follow-up probe	Example
General probe	How did you arrive at that answer? Tell me what you were thinking
Comprehension/ Interpretation probe	What does the term "health" mean to you?
Paraphrasing	Can you repeat the question I just asked in your own words?
Confidence judgement	How sure are you that you went to the doctor five times in the past 12 months?
Recall probe	How do you remember that you went to the doctor five times in the past 12 month

Gratitude: *First I would like to thank you for participating in the interview.*

Presentation of the study: *This study is being carried out as part of an international project in which…* (details about organizations involved)..

Objective of the study: *The objective of this study is to learn how people understand one of the questionnaires about…* (details about the questionnaire administered during the interview)

Objective of interview: *We are interested in knowing your opinion about …*(topic of the research) *and how you have interpreted the questions in the questionnaire, what you have thought about, what memories have come to mind when answering, etc.*

Confidentiality and use of personal data: *All responses you will give are confidential and the results will remain anonymous in reports that are made. Access to data is restricted to members of the research team and will only be used for the purposes of the study. The interview will be taped to facilitate analysis of the answers but sometimes I may take some notes during the course of the interview.*

Process: *First, I'll give you a booklet with the questionnaire and instructions on how to respond. Do not forget to read the instructions before starting to answer the questionnaire. In this first part of the interview, answer without asking questions. If you have questions, we can talk about them in the second part of the interview. When you have finished answering the questionnaire, we will begin the interview about how you answered the questionnaire…*

Asking for doubts: *Do you have any questions? Well let's start…*

Figure 2. *Relevant indications for respondent before the interview*

Data collection

In the context of validation studies of tests and questionnaires, the CI is usually done using a retrospective design (e.g., Willis, 2005). First, respondents respond to test or questionnaire items in conditions similar to those of their future application, and then begin the cognitive interview. The advantages and disadvantages of the retrospective design with alternative designs have been studied in the literature (e.g., Conrad & Blair, 2009). Cognitive interviews are often conducted in a "laboratory" equipped with audio and video recording devices. Those responsible for validation studies try to have experienced interviewers that are trained in the particular objectives of the study. The experience and training are key to the interviewers managing to access all the data on the test takers' psychological processes and cognitive operations, while not inducing their responses in the "rapport" of interviewer-interviewee interaction (Conrad & Blair, 2009).

After data collection, transcripts from recordings are obtained to facilitate the analysis process. Figure 3 shows an example of a transcript of an interview conducted in the study done by Benitez and Padilla (2013). The sequence of probes and answers was for the item "I will use science in many ways when I am an adult" included in the 2006 PISA Student Questionnaire (OECD, 2006).

Interviewer: where it says I will use science in many ways when I am an adult, what moments or situations were you thinking about?

Respondent: I was thinking about a career, or using science everyday

Interviewer: and you answered 'disagree' so tell me about your answer there

Respondent: I don't think I'm going to need know anything about science because I don't want a career in science, and most things like measuring, that's really easy, and I don't really see myself having a scientific career

Figure 3. *Example of probes and answer in transcriptions*

Analyses

Different approaches can be followed for analyzing data from cognitive interviews (Willis, 2005). The approach proposed by Willson and Miller (2013) is thought to be especially useful for conducting validation studies aimed at obtaining evidence about response processes. Figure 4 illustrates the main steps of the analytic process adapted to the testing context.

Analytic step	Tiers of theory building
1. Conducting	Individual response
2. Summarizing	Record of respondent difficulties Identification of potential themes
3. Comparing across respondents	Identification of "What the item captures"
4. Comparing across groups	Response process differences across groups
5. Concluding	Explanation of item performance

Figure 4. *Tiers of theory building for analytic steps*

Two aspects of the analytical process justify the recommendation: a) "Building Theory" means the development of the argument with which to examine the proposal that is the object of the validation study: the fit between psychological processes and cognitive operations of the test takers and those outlined in the test specifications, the similarity or difference between these processes across relevant groups of test takers etc. and b) the analytical process includes comparing the groups of respondents as defined by the validation study design.

Discussion

In this article we defined the source of evidence based on response processes, highlighted its relationship with the source of evidence based on the content of the test, identified when it is critical to obtain evidence based on response processes to justify the use of the test, and presented the methods available with special attention to the cognitive interview. The relationship between the evidence based on test content and those based on response processes is consistent with the unitary view of validity, and with the necessary accumulation of validity evidence to construct validity argument. In our opinion, this relationship has not been sufficiently exploited in the administration of validation studies despite its potential benefits in the study of bias, the testing of people with different linguistic antecedents, the adaptation of tests and questionnaires, or the testing of people with disabilities.

In parallel to the theoretical and conceptual work on the theory of validity, it is necessary to advance the development of innovative methods to facilitate the completion of the validation studies. Papers also aimed at optimizing the conditions of application of the methods are already available. For example, in the case of the cognitive interview method, it is necessary to address the problem of "reactivity" of the respondent, the design of follow-up probes which are reliable and efficient at capturing all the life experience that the "test takers" have when responding to test and questionnaire items, and translate to their answers.

It is also necessary that the field of Psychometrics contributes to breaking down the boundaries between quantitative and qualitative methodologies. Mixed research can make important contributions to the theory and practice of validation studies. The pillars of "pragmatism" and "integration" are perfectly consistent with the conceptual foundations of the current version of the validity theory.

With this article we hope to have contributed to researchers addressing validation studies based on response processes, provided that the validity argument being developed requires this type of validity evidence.

Acknowledgment

This study was partially funded by the Andalusia Regional Government under the Excellent Research Fund (Project nº SEJ-6569).

References

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education (1985). *Standards for educational and psychological testing*. Washington, D.C.: American Psychological Association.

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

Althof, S.E., Perelman, M.A., & Rosen, R.C. (2011). The Subjective Sexual Arousal Scale for Men (SSASM): Preliminary development and psychometric validation of a multidimensional measure of subjective male sexual arousal. *The Journal of Sexual Medicine*, *8*, 2255-2268.

Beatty, P., & Willis, G.B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly*, *71*, 287-311.

Benitez, I., & Padilla, J.L. (2013). Analysis of nonequivalent assessments across different linguistic groups using a mixed methods approach: Understanding the causes of differential item functioning by cognitive interviewing. *Journal of Mixed Methods Research*, Online first. doi: 10.1177/1558689813488245.

Borsboom, D., Mellenbergh, G.J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review, 111,* 1061-1071.

Brod, M., Hammer, M., Christensen, T., Lessard, S., & Bushnell, D.M. (2009). Understanding and assessing the impact of treatment in diabetes: The Treatment-Related Impact Measures for Diabetes and Devices (TRIM-Diabetes and TRIM-Diabetes Device). *Health and Quality of Life Outcomes*, *7*, 83.

Castillo, M., Padilla, J.L., Gomez, J., & Andres, A. (2010). A productivity map of cognitive pretest methods for improving survey questions. *Psicothema*, *22*, 475-481.

Castillo, M., & Padilla, J.L. (2012). How cognitive interviewing can provide validity evidence of the response processes to scale items. *Social Indicators Research*. Online first. doi: 10.1007/s11205-012-0184-8.

Cepeda, N.J., Blackwell, K.A., & Munakata, Y. (2013). Speed isn't everything: Complex processing speed measures mask individual differences and developmental changes in executive control. *Developmental Science*, *16*, 269-286.

Cizek, G.J., Rosenberg, S.L., & Koons, H.H. (2007). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement*, *68*, 397-412.

Cizek, G.J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods*, *17*, 31-43.

Collins, D. (2003). Pretesting survey instruments: An overview of cognitive methods. *Quality of Life Research, 12*, 229-238.

Conrad, F.G.; Blair, J. (2009). Sources of error in cognitive interviews. *Public Opinion Quarterly, 73*, 32-55.

Day, R.F. (2010). Examining the validity of the Needleman-Wunsch algorithm in identifying decision strategy with eye-movement data. *Decision Support Systems*, *49*, 396-403.

Deal, L.S., DiBenedetti, D.B., Williams, V.S., & Fehnel, S.E. (2010). The development and validation of the daily electronic Endometriosis Pain and Bleeding Diary. *Health and Quality of Life Outcomes*, *8*, 64.

Elling, S., Lentz, L., & de Jong, M. (2012). Combining concurrent think-aloud protocols and eye-tracking observations: An analysis of verbalizations and silences. *IEEE Transactions on Professional Communication*, *55*, 206-220.

Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*, 179-197.

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*, 24-35.

Ferdous, A.A., & Plake, B.S. (2005). Understanding factors that influence decisions of panelists in a standard setting study. *Applied Measurement in Education, 18,* 257-267.

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*, *100*, 37-60.

Garrett (1937). *Statistics in psychology and education*. New York: Longmans, Green.

Gehlbach, H., & Brinkworth, M.E. (2011). Measure twice, cut down error: A process for enhancing the validity of survey scales. *Review of General Psychology*, *15*, 380-387.

Hawthorne, G., Davidson, N., Quinn, K. McCrate, F., Winkler, I., Lucas, R., Kilian, R., & Molzahn, A. (2006). Issues in conducting cross-cultural research: implementation of an agreed international protocol designed by the WHOQOL Group for the conduct of focus groups eliciting the quality of life of older adults. *Quality of Life Research 15,* 1257-1270.

Ivie, J.L., & Embretson, S.E. (2010). Cognitive process modeling of spatial ability: The assembling objects task. *Intelligence*, *38*, 324-335.

Jabine, T.B., Straf, M.L., Tanur, J.M., & Tourangeau, R. (1984). *Cognitive aspects of survey methodology: Building a bridge between disciplines*. Washington, DC: National Academy Press.

Kane, M.T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*, 527-535.

Kane, M.T. (2006). Validation. In B.L. Robert (Ed.), *Educational Measurement* (4th ed., pp. 17-64). Wesport, CT: Praeger.

Kane, M.T. (2013). Validation as a pragmatic, scientific activity. *Journal of Educational Measurement*, *50*, 115-122.

Krall, J.S., & Lohse, B. (2010). Cognitive testing with female nutrition and education assistance program participants informs validity of the Satter eating competence inventory. *Journal of Nutrition Education and Behavior*, *42*, 277-283.

Krosnick, J.A. (1999). Survey research. *Annual Review of Psychology, 50*, 537-567.

Martin, E. (2004). Vignettes and respondent debriefing for questionnaire design and evaluation. En Presser, S., Rothgeb, J.M., Couper, M.P., Lessler, J.T., Martin, E., Martin, J., & Singer, E. (2004). *Methods for Testing and Evaluating Survey Questionnaires*. (pp. 149-173). New York: Wiley-Interscience.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational Measurement 3rd ed*. (pp. 13-103). New York: McMillan.

Messick, S. (1990). *Validity of test interpretation and use*. Research Report 90-11. Education Testing Service.

Miller, K., Chepp, V., Willson, S., & Padilla, J.L. (2014 in press). *Cognitive Interviewing Methodology: A Sociological Approach for Survey Question Evaluation*. New York, NJ: John Wiley and Sons.

Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (2002). On the role of task model variables in assessment design. In S.H. Irvine & P.C. Kyllonen (Eds.), *Item generation for test development* (pp. 97-128). Mahwah, NJ: Lawrence Erlbaum.

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 Professionals-Revised (IAPCC-R). *Journal of transcultural nursing: Official journal of the Transcultural Nursing Society / Transcultural Nursing Society*, *21*, 55-64.

Sireci, S.G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R.W. Lissitz (Ed.), *The Concept of Validity* (pp. 19-39). Charlotte, NC: Information Age Publishing, Inc.

Sireci, S.G. (2012). "De-constructing" Test Validation. Paper presented at the annual conference of the National Council on Measurement in Education as part of the symposium "Beyond Consensus: The Changing Face of Validity" (P. Newton, Chair). April 14, 2012, Vancouver, Canada.

Sireci, S., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema, 26*, **xx-xx???.**

Sireci, S.G., & Parker, P. (2006). Validity on trial: Psychometric and legal conceptualizations of validity. *Educational Measurement*, *25*, 27-34.

Sireci, S.G., Baldwin, P., Martone, A., Zenisky, A.L., Kaira, L., Lam, W., Shea, C.L., Han, K., Deng, N., Delton, J., & Hambleton, R.K. (2008). *Massachusetts adult proficiency tests technical manual: Version 2*: Amherst, MA: Center for Educational Assessment. Available at http://www.umass.edu/remp/CEA_TechMan.html.

Skorupski, W.P., & Hambleton, R.K. (2005).What are panelists thinking when they participate in standard setting studies? *Applied Measurement in Education, 18,* 233-256.

Tourangeau, R. (1984). Cognitive science and survey methods: A cognitive perspective. En T. Jabine, M. Straf, J. Tanur & R. Tourangeau (Eds.), *Cognitive Aspects of Survey Methodology: Building a Bridge Between the Disciplines* (pp. 73-100). Washington, DC: National Academy Press.

Tourangeau, R., Rips, L.J., & Rasinski, K. (2004). *The Psychology of Survey Response*. Cambridge: Cambridge University Press.

Wang, X., & Sireci, S.G. (2013). Evaluating the cognitive levels measured by test items using item response time. Paper presented at the annual conference of the American Educational Research Association. April, 2013, San Francisco, USA.

Webber, M.P., & Huxley, P.J. (2007). Measuring access to social capital: The validity and reliability of the Resource Generator-UK and its association with common mental disorder. *Social Science & Medicine (1982)*, *65*, 481-492.

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

Willson, S., & Miller, K. (2014 in press). Data collection. In K. Miller, V. Chepp, S. Willson & J.L. Padilla (Eds.), *Cognitive Interviewing Methodology: A Sociological Approach for Survey Question Evaluation*. New York, NJ: John Wiley and Sons.

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* (pp. 65-83). Charlotte, NC: Information Age Publishing, Inc.

Zumbo, B.D., & Shear, B.R. (2011). The concept of validity and some novel validation methods. In *Northeastern Educational Research Association* (p. 56). Rocky Hill, CT.