Paidéia
2023, Vol. 33, e3336.doi:https://doi.org/10.1590/1982-4327e3336
ISSN 1982-4327 (online version)

Thematic Dossier "Methodological Challenges in Psychology: Contributions to Academic Practice and Training"

Cognitive Interview in the Search for Validity Evidence Based on Response Processes

Ana Cláudia Araújo da Cruz[1]
Ana Gabriela Rocha Araújo[1]
Cláudio Simon Hutz[1]
Cleonice Alves Bosa[1]

Abstract: Cognitive interviews can provide validity evidence for instruments based on item response processes; however, use of focus groups still prevails in Brazilian literature. Moreover, semantic analysis has only been considered when searching for validity evidence based on test content. This paper presents a proposal for qualitative data analysis based on cognitive interviewing, thus providing researchers with a protocol that enables best practices in carrying out this technique, and consolidating it in the Brazilian literature as an option to search for validity evidence based on item response processes. To conclude, we present some criticisms regarding current procedures for validity evidence based on test content and discuss some possibilities.

Keywords: item content (test), test construction, test validity, psychological testing

Entrevistas Cognitivas na Busca de Evidências de Validade Baseadas no Processo de Resposta

Resumo: As entrevistas cognitivas podem fornecer evidências de validade para os instrumentos com base no processo de resposta aos itens. Na literatura brasileira, entretanto, o uso de grupos focais ainda prevalece. Além disso, a análise semântica tem sido considerada apenas na busca de evidências de validade baseada no conteúdo do teste. Esse manuscrito apresenta uma proposta para análise de dados qualitativos de entrevistas cognitivas. O objetivo é fornecer aos pesquisadores um protocolo que viabiliza as melhores práticas na realização desta técnica, consolidando-a na literatura brasileira como uma opção para busca de evidência de validade baseada no processo de resposta dos itens. Por fim, são apresentadas algumas críticas em relação aos atuais procedimentos de busca de evidências de validade baseadas no conteúdo e possibilidades são discutidas.

Palavras-chave: conteúdo do item (teste), construção do teste, validade do teste, testes psicológicos

Entrevistas Cognitivas en la Búsqueda de Evidencia de Validez Basada en el Proceso de Respuesta

Resumen: Las entrevistas cognitivas pueden proporcionar evidencia de validez para los instrumentos basados en el proceso de respuesta al ítem. En la literatura brasileña, sin embargo, aún prevalece el uso de grupos focales; además, el análisis semántico solo se ha considerado en la búsqueda de evidencias de validez basadas en el contenido de la prueba. Este manuscrito presenta una propuesta para el análisis de datos cualitativos de entrevistas cognitivas. El objetivo es proporcionar a los investigadores un protocolo que posibilite las mejores prácticas en la realización de esta técnica, consolidándola en la literatura brasileña como una opción para la búsqueda de evidencias de validez a partir del proceso de respuesta a los ítems. Finalmente, se presentan algunas críticas en relación a los procedimientos actuales de búsqueda de evidencias de validez en base al contenido y se discuten posibilidades.

Palabras clave: contenido del item (test), construcción de test, validacion de test, tests psicológicos

Every year, the number of studies using data collected from self-report scales increases considerably (Fryer & Nakao, 2020); however, the use of this type of instrument has been highly criticized. In 1946, Ellis conducted an extensive literature review that included validation studies of specific self-report instruments for measuring personality. In his conclusions, the author stated that the use of these instruments was "of dubious value in distinguishing between groups of adjusted and maladjusted individuals,

[1]*Universidade Federal do Rio Grande do Sul, Porto Alegre-RS, Brazil*

Article derived from the dissertation by the first author under the supervision of the third, defended in 2021, in the Postgraduate Program in Psychology at the Universidade Federal do Rio Grande do Sul. Support: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Financing Code 001- and the National Council for Scientific and Technological Development - CNPq).

Correspondence address: Ana Cláudia Araújo da Cruz. Universidade Federal do Rio Grande do Sul. Rua Ramiro Barcelos 2600, Porto Alegre-RS, Brazil. CEP 90.035-003. E-mail: anaclaudiacruz.psico@gmail.com

and that they are of much less value in the diagnosis of individual adjustment or personality traits" (Ellis, 1946, p. 426). Developments in the understanding of constructs and advances in statistical methods have rendered most of Ellis' considerations obsolete. Currently, few researchers are able to avoid using this type of instrument in social sciences research (Fryer & Nakao, 2020) and self-reporting has come to be considered the most appropriate, or even the only appropriate method for measuring some psychological phenomena (Clark & Watson, 2019). However, one of the issues identified by Ellis still remains relevant: different individuals can interpret the same items in different ways (Ellis, 1946); and, possibly for this reason, developing suitable self-report items for measuring psychological constructs remains a major challenge faced by researchers.

Suitability, in terms of how items are understood, has been tested using semantic analysis. In Brazil, the use of semantic analysis is well documented in the literature. Pasquali (1998) points out that the process must consider two objectives: to ensure their intelligibility to the less educated and to prevent them from being inelegant to the more educated population. Pacico (2015) states the need for the items to be understood in the same way by respondents, but points out that the semantic analysis aims to assess "how understandable and clear the items are and whether they have apparent validity" (p. 64). According to Borsa and Seize (2017) the semantic analysis must ensure that the items are properly understood, proposing that they be presented orally to the participants, who must reproduce them and explain what they understood. Most of the analysis proposals presented by these authors include the relations between the content of the instrument and the construct it aims to measure, i.e., they provide validity evidence based on the test's content (American Educational Research Association [AERA] et al., 2014). The procedures suggested for analyzing items by national authors, which generally include conducting focus groups, are limited to the premise that the response to a self-report instrument depends on the respondents' understanding of the task, reading comprehension and level of vocabulary (Urbina, 2014). However, the complexity involved in the process goes far beyond this understanding.

Tourangeau (1984) proposed a model to broaden the understanding of the response process, considering four stages: (a) comprehension; (b) retrieval; (c) judgment; and d) response selection. The stages influence each other systematically, errors can occur at any of the four stages and not all stages are always completed or follow the same order, although most of the time this is the case (Tourangeau & Hanover, 2018). The first stage, in which the interviewee must interpret the item and understand its intention, includes several operations. The process begins with the interviewee bringing up the sounds of words in their working memory and dividing these words into groups that will form meaningful concepts. If there are different possible interpretations, they should select one of them, which can be reviewed later to check for plausibility. By establishing relations between the concepts formed, the interviewees construct the meaning of the entire sentence. They then abandon the original words and keep only their interpretation when moving on to the second stage. At this stage, the interviewees begin to retrieve memories that include relevant events, thoughts and beliefs, guided by their understanding of the item. But retrieving a piece of information or belief does not guarantee that it will be used in judging the item. Interviewees may suppress information if they consider it suspicious or of little relevance, since the judgment stage requires several pieces of information to be combined (Tourangeau, 1984). The last stage is selecting the response. In cases where Likert-type scales are used, the semantic analysis should first ensure that item (in direction and degree of intensity) and scale comprehension are coherent. Moreover, interviewers should be aware of social desirability effects, in cases where the retrieved memories do not match the response selected (Tourangeau & Hanover, 2018). Thus, investigation by semantic analysis should not be limited to answering whether the interviewee understood the question; rather, it should ensure that the comprehension reached through the whole process is free of potential sources of confusion and corresponds to the researcher's understanding. In other words, that the interviewee's comprehension actually reaches the same domain, the same evaluative dimension for which the item was created (Hubley, 2021; Schwarz, 1999).

In its most recent edition, the Standards for Educational and Psychological Testing (AERA et al., 2014) stresses the importance of the validation process including the search for validity based on the response process. Although they do not clearly state which methods should be used in this search, they emphasize the analysis of individual responses from part of the public for whom the instrument is intended. Searching for this type of evidence must consider both the respondents and the data, which hinders developing suitable methods (Padilla & Benítez, 2014). Search for validity evidence based on the response process is possibly one of the most complex stages in the construction or adaptation of an instrument (Hubley, 2021), and only recently has this stage been officially supported by Cognitive Interview (CI) (Peterson et al., 2017).

CI was developed in the 1980s, founded on cognitive psychology, due to the need to understand the mental processing that takes place while people are answering questionnaires (Willson & Miller, 2014). The technique involves an in-depth investigation, which aims to understand what goes through participants' minds when they answer questions or items (Willis, 2005). CI should therefore be understood as a process made up of distinct stages, including a clear identification of the item's intention, data collection and analysis, and comparison of the results obtained with the expected results (Peterson et al., 2017; Wolcott & Lobczowski, 2021). Two strategies are generally considered for conducting CI: thinking aloud and verbal probing. When subjected to the think-aloud strategy, participants are instructed to verbalize the thoughts that occur while answering the items; the interviewer only

encourages the participant, avoiding interference as much as possible (Willson & Miller, 2014). Verbal probing, on the other hand, is characterized by an active process of investigation by the researcher based on pre-defined questions that guide the interviewee in verbalizing mental processes (Willis, 2005). Despite a clear difference established in the literature between the think-aloud and verbal probing strategies, especially regarding the role of the interviewer, they are most commonly used in combination. Thus, the interviewee is instructed to think aloud, but the interviewer asks questions when necessary to encourage the respondent to carry out this process (Meadows, 2021). The frequent use of both strategies is possibly due to the poor performance shown by many participants when subjected to the think-aloud process (Willis, 2018).

Although one cannot simply generalize the results obtained through CI, since it is impossible to judge how common the problems pointed out by the participants are, the method provides important information about how interviewees actually formulate their responses and about the types of errors that occur during this process that other methods would not reveal (Tourangeau et al., 2019).

Peterson et al. (2017) proposed the use of CI as a strategy to search for validity evidence based on the response process, for self-report instruments (constructed or adapted). Their proposal seeks to help identify sources of confusion at each stage of the process. However, the authors do not detail how the process should be systematized so that researchers can avoid errors that compromise data collection. Moreover, there is no clear description of how the constitutive and operational definitions of a construct can favor data collection and analysis.

To date, no Brazilian publication has examined how semantic analysis can be used to evaluate response processes. When compared to other validation methods, studies presenting this category of evidence are scarce, which may be due precisely to the lack of clearer protocols for its realization (Hubley, 2021; Padilla & Benítez, 2014). Our proposal intends to fill this gap, and is illustrated by situations that occurred during the construction of an instrument based on the Big Five personality model. Its elaboration considers both the notes of major CI researchers and elements from the present authors' experiences. Our present conclusions were reached by conducting the CI process and considering the mistakes made and successes obtained.

This is a methodological article (American Psychological Association [APA], 2020), since it proposes a method for analyzing qualitative data collected by cognitive interviews, in which empirical data were used only for illustration. The process was thoroughly discussed by the working group to reach a consensus on the best practices for carrying out the CI technique as a strategy to search for validity evidence based on the response process. We aim to provide researchers with a protocol that enables best practices in conducting Cognitive Interview processes, consolidating it in the Brazilian literature as a method of searching for validity evidence based on response processes.

Cognitive Interviews in Practice

Previous steps: Preparing for the interviews

To construct an instrument, one must begin by defining the domain to be measured (Boateng et al., 2018), which includes a constitutive and operational definition of the construct. In other words, before constructing the items, the construct needs to be conceptualized in detail using existing literature. Constitutive definition is very similar to a dictionary entry, in which a concept is defined by other concepts, whereas the operational definition must be offered on the basis of concrete operations and behaviors (Pasquali, 1998). Although we do not intend to cover the entire process of constructing an instrument, its definition is being addressed here because it will guide our analysis of the interviews. From these two definitions it will become clear to the researcher what Peterson et al. (2017) call 'item intent.' With a clear view of the domain represented by the item, we can determine whether the content evoked by the participant is truly representative of the construct. In this regard, the first stage involves drafting a document containing the constitutive definition, the operational definition and the assessment items. This document will later be used by experts to evaluate the data collected by means of interviews, but it will also guide the interviewer to probe the answers provided in order to clarify possible ambiguities.

A second document that will be used during CI must also be drawn up beforehand. The items must be arranged in an organized manner within a protocol that will be given to the respondent. To construct this document, one should consider that interviewees will be asked to read (aloud) the first item, explain what they understood, verbalize the thoughts that occurred during the response process, and return to the protocol to read the next item. It is therefore important that the document identify the items in some way, such as numbers. Moreover, for constructs that have more than one factor or factors composed of facets, varied protocols that do not present items from the same domain in sequence are suggested. The participant should also receive a guide containing the response options. Usually, self-reporting instruments use Likert-type scales, and it based on this guide that participants can judge the item and choose the answer.

Tourangeau (1984) discusses different ways of presenting the item, including reading by the interviewer (oral presentation) and reading by the interviewee (written presentation). We suggest that the item be presented in written form and that the interviewee be asked to read it aloud. This presentation format was chosen because it can provide important information about item fluency, and replicate the way in which the instrument will be answered once finished. Use of protocols that contemplate a different order of items aims to minimize the impact that immediately preceding items may have on the following items (Schwarz, 1999). Non-sequential arrangement of items from the same dimension seeks to ensure that item comprehension has not been affected (facilitated) by previous items, since items from the same domain can present very similar retrieved content.

The interviewees: Selecting the sample

This is possibly the topic that will require the most discussion by researchers. Not just because the literature is unclear on the subject. Answers to who should make up the sample and how many interviews are necessary depend exclusively on the characteristics and complexity of the construct and the instrument being assessed (Willson & Miller, 2014). Participant selection must consider, above all, the population for which the instrument is intended. Ideally, these participants should not be part of a restricted section of the population, such as university students (unless the instrument is intended so). Pasquali (1998) suggests considering two distinct poles: individuals with low and high schooling levels. According to the author, individuals from the lowest (ability) stratum of the target population provide information on how the items are being read and understood. On the other hand, the higher education level would help to avoid inelegance in item wording Although this choice can provide relevant information, since the answers depend on the respondents' cognitive processing (Urbina, 2014), CI's main interest is to know if completely different individuals respond to the items with equal mastery. If the proposal is to create an instrument to investigate work-related engagement, it makes more sense to consider individuals from different functions and professions, rather than selecting the sample solely based on schooling.

There is no a priori exact number of CIs suitable for evaluating self-report items, although they generally range from five to fifteen interviews (Peterson et al., 2017). Considering the objective of identifying possibly problematic items, the sample size must be sufficient to ensure that problems have been identified (Willis, 2005). More important than the number of interviewees, however, is the right choice of sample, the proper conduct of the interviews and a careful analysis of the results obtained. The number of interviews cannot be decided based on a set goal; the decision must ponder the extent to which the data collected was able to contribute to the objective of the process carried out (Willson & Miller, 2014). If an item has been evaluated by three participants, resulting in three different understandings, the most appropriate course of action is to change the item based on the data obtained before submitting it to new interviews; or to exclude it, rather than continuing to conduct interviews that will only corroborate the item's already known inadequacy.

The interview: Collecting data

Before starting the interviews, we must ensure that the participant has been sufficiently informed about the purpose of the research. Emphasize that it is the items that are under analysis, that is, what is being assessed is the ability of the items to evoke satisfying responses to the researcher's quest, and not the participant's ability to respond adequately to them. Hence, there are no right or wrong answers. Moreover, no information about the construct being measured should be provided to prevent participants from forming a prior impression of the items.

A standard text may be used, such as: "I am going to give you a sheet of paper with a few sentences and we need to evaluate each one. To do so, I need you to read the first sentence out loud and tell me what you understand, that is, what it means. When you have completed this first part, in the first sentence you will need to tell me how much it represents you by ticking the corresponding number on the scale below, in which 1 means it does not represent you at all, it has nothing to do with you; and (maximum number of categories on the scale) represents you a lot, it has a lot to do with you. But there is an important detail: during this process you will describe to me what you are thinking, remembering, in other words, what is guiding you in your choice. When we have finished this part, we will move on to the second sentence and start the same process. There is no right or wrong answer, the most important thing is that you describe your thoughts to me in as much detail as possible."

At first, the technique may seem strange to the interviewee and practicing thinking aloud can help them understand the task. Willis suggests a training session in which the interviewee is given the following orientation: "visualize the place where you live, and think about how many windows there are in that place. As you count up the windows, tell me what you are seeing and thinking about" (Willis, 2005). Researchers can also create their own training vignettes, as long as they are unrelated to the construct underpinning the instrument to be evaluated; or use one or two initial items that are not part of the instrument and which allow the researcher to make more targeted scores, without compromising protocol validity.

Data collection should preferably be conducted using a tape recorder, which will allow the researcher greater freedom to intervene verbally, without having to take notes. The think-aloud technique should be the first option; however, many interviewees need encouragement to be able to accomplish the task (Willis, 2018). In this regard, the most important thing is that the researcher has been prepared for the task and is able to conduct the verbal survey properly (Wolcott & Lobczowski, 2021). Their scores should be neutral, with the sole aim of facilitating the interviewee's verbalization and clarifying specific doubts. It is not possible to develop an intervention script in advance, because the sources of confusion cannot be anticipated (Willis, 2005). The researcher must therefore be prepared to probe the interviewee with as little involvement as possible. Questions such as: "What are you thinking?" or "I can see that you were in doubt, what went through your mind?" can be a good alternative to encourage verbalization. However, researchers must be careful that the direction does not lead the interviewees to simply justify their answers, rather than describing the response process (Peterson et al., 2017).

Figure 1 shows examples of questions that can help clarify the four stages of the response process. This survey model is based on notes by Tourangeau (1984) and Willis (2005), and is a suggestion. As can be seen in the examples used later,

interviewers with greater expertise are able to answer questions using different resources. The most important thing at this stage is for the interviewer to understand their responsibility to identify and explore inconsistencies in the data relating to each response process step, and to be prepared to ask questions capable of clarifying any existing doubts (Willson & Miller, 2014).

Figure 1

Examples of questions considering the response process stages

Possible questions to guide each stage response process step	
1) Comprehension	• What does X mean to you?
2) Retrieval	• What are you remembering now? • Is there anything specific that the item reminded you of?
3) Judgment	• What made you choose that response? • What made you think this response describes you best?
4) Response selection	• Was it easy to choose that response? • Do you think that response describes you well?

Note: Source: Prepared by the authors.

The results: Analyzing the data

There is also no consensus in the literature on how CI data should be analyzed. A method widely used in qualitative research includes synthesizing and reducing the content by creating categories (Miller et al., 2014). Although this procedure can be very useful in the construct's constitutive and operational definition phase, in searching for validity evidence based on response processes it may not be very useful. This is because validation does not intend to generate categories that seek to clarify the construct, but to identify whether the construct (already established) is being achieved through the items (already elaborated).

The first analysis should be conducted by the researcher during the interview. This needs to be stressed, because interviewers must be attentive to the answers in order to identify whether they met expectations. In other words, if they provided enough information and did not deviate from the proposed topics (Miller et al., 2014). This will allow the interviewer to conduct the survey process in such a way as to ensure that relevant information has been obtained. The second stage of data analysis should be conducted by having the items checked individually by independent experts. Unlike what was done for the interviews, the items should be regrouped considering their original domain. This will facilitate the experts' job and prevent them from having to return to the constitutive and operational definition of the domain for each item assessed. Data can be presented in two ways: the expert can have access to the transcribed material, or to the interview recording itself (as long as it is organized in such a way as to guarantee evaluation fluidity). Figure 2 shows an example of how the material to be given to experts could be drawn up.

The following are examples of how the proposed protocol can help in item analysis. Figure 3 shows the process performed with an item from the Warmth facet belonging to the Extroversion factor.

Based on the operational definition of the construct, the researchers identified that highly warm people tend to tell others more about how important they are and how happy they are to have them around. Despite being a striking characteristic of highly warm individuals, the creation of an item considering exclusively verbal behavior related to this dimension caused a measurement error. From the interview, we identified that in the retrieval process, the interviewee separated verbal behaviors from other behaviors related to the domain, and provided a false score. In this example, the scoring conducted by the researcher was not a question and was naturally included into the context of the interview to confirm that the score provided by the participant did not represent a real score for the dimension being assessed. Figure 4 shows another example of an item from the Extroversion factor, this time considering the Excitement-Seeking facet.

Considering the need for stimulation presented by individuals with a high degree of Excitement-Seeking, the item was constructed to measure individuals who have a high score on the facet and are unable to be satisfied with the present stimuli. In this case, the maximum score would reflect the individual's high level. However, the interviews revealed that although the comprehension, retrieval and judgment steps meet expectations, a problem with the response scale means that the item does not adequately measure the latent trait. If the respondent scores high, it may be because they need stimulation. However, if they score low, they may do so because they do not need these stimuli (low Excitement-Seeking score), or because they need them but already have them in a satisfactory quantity (high Excitement-Seeking score).

Although the examples presented here have been chosen precisely because they illustrate important errors (which signal that item validity is already compromised during data collection), smaller errors can go unnoticed at first and are only pointed out by the experts' analysis. Some situations may still require arbitrary choices on the part of the researcher. We noticed during data collection that some peculiar interpretations were reported, leading to discussions about when errors in the response process were actually caused by item structure, rather than by the individual's singular experience. In cases where peculiar characteristics are observed, the experts will be able to help researchers in their decision.

Finally, the indices suggested in the literature for quantitative evaluation of the expert judges' analysis, such as the agreement percentage, the Content Validity Index and the Kappa Coefficient (Borsa & Seize, 2017), can also be used in the search for evidence based on response processes. An item that has undergone three CI and has been evaluated by three experts, for example, will provide nine indicators that can be evaluated using methods that quantify the degree of agreement between experts. These methods are already well documented and are usually used to find validity evidence based on test content (Almanasreh et al., 2019).

Figure 2

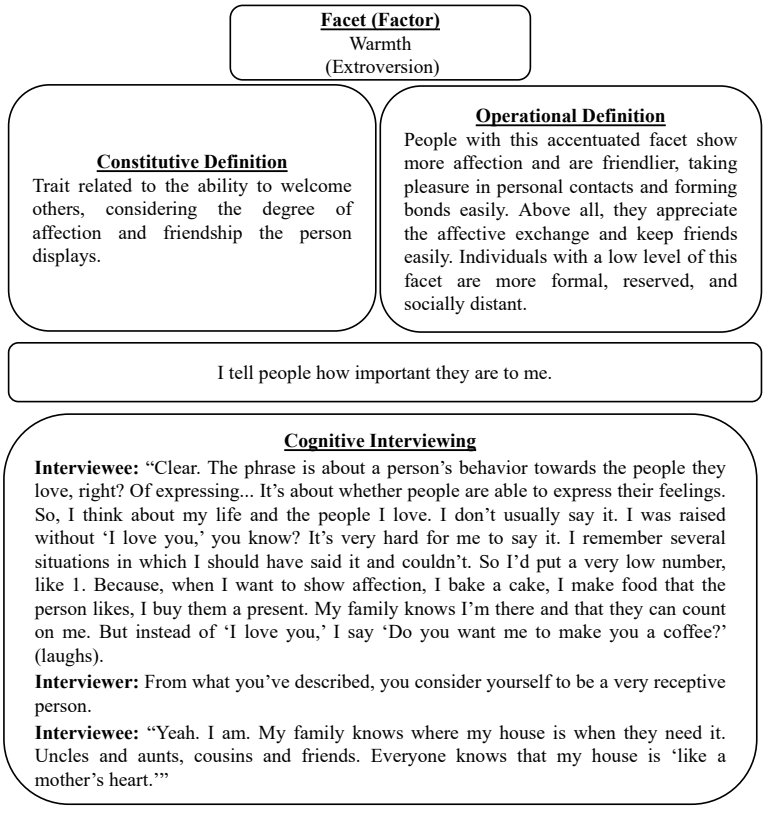
Example protocol for evaluating experts

Dimension: Constitutive Definition: Operational Definition:				
Item 1:	Poor	Fair	Good	Great
Does the initial reading of the item show satisfactory fluency? (Do not use for transcribed interviews)				
Is the initial comprehension of the item as expected?				
During the retrieval process, did the participant present content that was relevant to the domain being assessed (and only to that domain)?				
When judging the response options, did the participant stick to the content retrieved?				
Is the response chosen in line with the process conducted by the participant?				

Note: Source: Prepared by the authors.

Figure 3

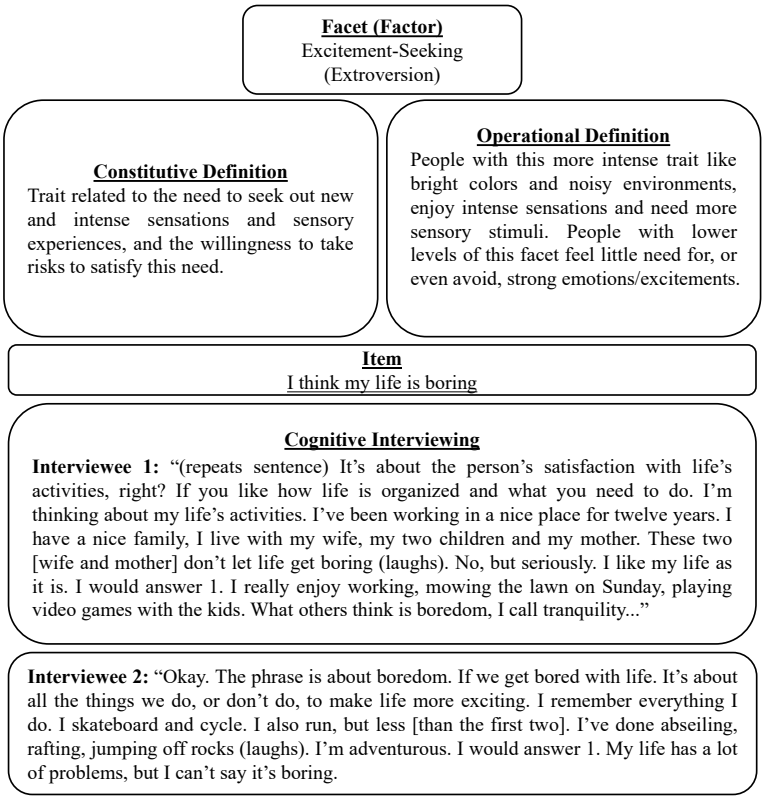
Example of item evaluated - Warmth facet / Extroversion factor



Note: Source: Prepared by the authors.

Figure 4

Example of item evaluated - Excitement-Seeking facet / Extroversion factor



Note: Source: Prepared by the authors.

Validity Based on Test Content:

Current scenario and possibilities

The Brazilian literature is clear on the steps to be followed when searching for content-based validity evidence for self-report scales: content analysis conducted by experts, followed by semantic analysis conducted by the target population (Borsa & Seize, 2017; Pacico, 2015; Pasquali, 1998). However, we identified several problems with the practical application of this process. The first concerns an important divergence between some expert evaluations, which has already been addressed in the literature (Tourangeau et al., 2019). We asked each expert judge to score the items on a four-point scale (1=Bad, 2= Fair, 3= Good and 4= Great) in two domains: verbal comprehension and fluency, which we call Clarity; and relevance and representativeness, which we call Relevance. The results revealed several items scored at opposite ends by different judges. While one expert signaled that the item was bad in one or both domains, the other signaled that it was great. This may be because, when judging item relevancy, the expert considers their own response process. When content comprehension and retrieval are equivalent to what is expected, the judge considers the item to be adequate; conversely, when comprehension and retrieval are different from what is expected, the judge considers the item to be inadequate. The second problem highlighted concerns continuity. Many items scored as good or excellent by the experts turned out to be problematic during semantic analysis, proving inadequate for measuring the construct. The necessary changes made to the items after semantic analysis sometimes made them completely different from the items initially evaluated, raising questions about the initial content validity attested to.

Peterson et al. (2017) suggest that CI can also provide validity evidence based on test content. Thus, at the end of item evaluation, the interviewee should be asked about the construct. According to the authors, the researcher should explain to interviewees the domain being measured and question them about item suitability. This procedure would also allow to assess construct coverage, functioning as a validity evidence based on test content. In our experience, the results obtained with this procedure added no relevant information to the process. The participants were unable to provide information that went beyond that already presented during item evaluation. This observation is corroborated by Willson and Miller's (2014) comment that the respondent is an expert only in their personal experiences and not in the construction of items or instruments, and should not be asked to perform this function. However, we believe that CI can play a supporting role when this same question is put to the expert judges, indirectly contributing to the search for evidence of test content validity.

Given these findings, we suggest that the expert analysis stage in the search for content validity evidence be carried out after the semantic analysis, contrary to what is currently proposed in the literature. Instead, we propose that it be included at the end of the CI data evaluation. This procedure converges with that proposed by the Standards for Educational and Psychological Testing (AERA et al., 2014), which emphasizes that the search for validity by response process can include observers or experts invited to evaluate the data considering construct interpretation and definition; and is also in line with the concept of complementarity between sources proposed by Padilla and Benítez (2014). We believe that immersing ourselves in the constitutive and operational definition of the construct and in the response process described by interviewees can favor the analysis of each item and domain in its entirety, avoiding such marked divergences and favoring the obtaining of validity evidence based on test content.

If the researcher chooses to obtain validity evidence based on content considering the proposed model, fields that allow experts to assess item relevance and representativeness should be included at the end of the protocol presented. In this case, the experts should be explicitly informed about their role: first assess the response process presented; and then, considering the content presented, evaluate the relationship between the content of the instrument and the construct it is intended to measure. Further studies, considering the results obtained and the experts' own perception of the process, could confirm or refute the viability of this process.

Final Considerations

Although CI is being widely used for test development, especially internationally, there are still major variations in how procedures are conducted and reported (Meadows, 2021). The study sought to fill an existing gap in the national literature regarding CI practices to search for validity evidence based on the response process.

In addition to providing a protocol based on Tourangeau's (1984) response model to qualify this search, we also question the current practice in which items are first submitted to an expert evaluation and only then to semantic analysis. This is because the semantic analysis stage can lead to changes in the items writing, compromising the evidence of content validity achieved in the initial process. In this study, we propose a reversal of these steps, in which the experts are asked to evaluate the items after also having access to the CI data. We believe that providing experts with the constitutive and operational definition of the item, the item itself and the response-related content reported by the interviewees, can favor the analysis performed, qualifying the process of obtaining both validity evidence based on test content and validity evidence based on the response process.

Like the other techniques considered in the validation process, CI has some limitations. We cannot say, for example, whether the same results would be achieved by different researchers; nor is it clear how large a sample is needed for data saturation (Willis, 2018). In this regard, the technique still needs further studies. Besides, this manuscript is not intended to be a one-size-fits-all guide for obtaining this type of validity evidence. We only hope to contribute to researchers seeking to qualify the process of validating self-reporting instruments, providing a structured alternative and encouraging discussion about the use of qualitative methods to construct measurement instruments.

References

Almanasreh, E., Moles, R., & Chen, T. F. (2019). Evaluation of methods used for estimating content validity. *Research in Social and Administrative Pharmacy, 15*(2), 214-221. https://doi.org/10.1016/j.sapharm.2018.03.066

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). The s*tandards for educational and psychological testing.* AERA.

American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th ed.). APA.

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health, 6*, 149. https://doi.org/10.3389/fpubh.2018.00149

Borsa, J. C., & Seize, M. M. (2017). Construção e adaptação de instrumentos psicológicos: Dois caminhos possíveis [Construction and adaptation of psychological instruments: Two possible paths]. In B. Damasio & J. C. Borsa (Eds.), *Manual de desenvolvimento de instrumentos psicológicos* [Psychological instrument development manual] (pp. 15-38). Vetor.

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment, 31*(12), 1412-1427. https://doi.org/10.1037/pas000626

Ellis, A. (1946). The validity of personality questionnaires. *Psychological Bulletin, 43*(5), 385-440. https://doi.org/10.1037/h0055483

Fryer, L. K., & Nakao, K. (2020). The future of survey self-report: An experiment contrasting likert, VAS, slide, and swipe touch interfaces. *Frontline Learning Research, 8*(3), 10-25. https://doi.org/10.14786/flr.v8i3.501

Hubley, A. M. (2021). Response processes validity evidence : Understanding the meaning of scores from psychological measures. In P. Graf & D. J. A. Dozois (Eds.), *Handbook on the state of the art in applied psychology* (pp. 413-434). Wiley Blackwell.

Meadows, K. (2021). Cognitive interviewing methodologies. *Clinical Nursing Research, 30*(4), 375-379. https://doi.org/10.1177/10547738211014099

Miller, K., Willson, S., Chepp, V., & Ryan, J. M. (2014). Analysis. In K. Miller, S. Willson, V. Chepp, & J. L. Padilla (Eds.), *Cognitive interviewing methodology* (pp. 35-50). Wiley.

Pacico, J. C. (2015). Como é feito um teste? Produção de itens [How is a test built? Item production]. In C. S. Hutz, D. R. Bandeira, & C. M. Trentini (Orgs.), *Psicometria* [Psychometrics] (pp. 55-70). Artmed.

Padilla, J.-L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema, 26*(1), 136-144. https://doi.org/10.7334/psicothema2013.259

Pasquali, L. (1998). Princípios de elaboração de escalas psicológicas [Principles of elaboration of psychological scales]. *Revista de Psiquiatria Clínica, 25*(5), 206-213. http://ppget.ifam.edu.br/wp-content/uploads/2017/12/Principios-de-elaboracao-de-escalas-psicologicas.pdf

Peterson, C. H., Peterson, N. A., & Powell, K. G. (2017). Cognitive interviewing for item development: Validity evidence based on content and response processes. *Measurement and Evaluation in Counseling and Development, 50*(4), 217-223. https://doi.org/10.1080/07481756.2017.1339564

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54*(2), 93-105. https://doi.org/10.1037//0003-066x.54.2.93

Tourangeau, R. (1984). Cognitive sciences and survey methods. In T. B. Jabine, M. L. Straf, J. M. Tanur, & R. Tourangeau (Eds.), *Cognitive aspects of survey methodology: Building a bridge between disciplines* (pp. 73-100). The National Academies.

Tourangeau, R., & Hanover, L. (2018). The survey response process from a cognitive viewpoint. *Quality Assurance in Education, 26*(2), 169-181. https://doi.org/10.1108/qae-06-2017-0034

Tourangeau, R., Maitland, A., Steiger, D., & Yan, T. (2019). A framework for making decisions about question evaluation methods. In P. Beatty, D. Collins, L. Kaye, J. L. Padilla, G. Willis, & A. Wilmot (Eds.), *Advances in questionnaire design, development, evaluation and testing* (pp. 47-73). Wiley.

Urbina, S. (2014). *Essentials of psychological testing* (2nd ed.). Wiley.

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. SAGE.

Willis, G. B. (2018). Cognitive interviewing in survey design: State of the science and future directions. In D. Vannette & J. Krosnick (Eds.), *The Palgrave handbook of survey research* (pp. 103-107). Palgrave Macmillan.

Willson, S., & Miller, K. (2014). Data collection. In K. Miller, V. Chepp, S. Willson, & J. L. Padilla (Eds.), *Cognitive interviewing methodology* (pp. 15-33). Wiley.

Wolcott, M. D., & Lobczowski, N. G. (2021). Using cognitive interviews and think-aloud protocols to understand thought processes. *Currents in Pharmacy Teaching & Learning, 13*(2), 181-188. https://doi.org/10.1016/J.cptl.2020.09.005

Ana Cláudia Araújo da Cruz is a Ph.D. candidate of the Instituto de Psicologia at the Universidade Federal do Rio Grande do Sul, Porto Alegre-RS, Brazil.

Ana Gabriela Rocha Araújo is a Ph.D. candidate of the Instituto de Psicologia at the Universidade Federal do Rio Grande do Sul, Porto Alegre-RS, Brazil.

Cláudio Simon Hutz is a Professor of the Instituto de Psicologia at the Universidade Federal do Rio Grande do Sul, Porto Alegre-RS, Brazil.

Cleonice Alves Bosa is a Professor of the Instituto de Psicologia at the Universidade Federal do Rio Grande do Sul, Porto Alegre-RS, Brazil.

Authors' Contribution:
All authors made substantial contributions to the conception and design of this study, to data analysis and interpretation, and to the manuscript revision and approval of the final version. All the authors assume public responsibility for content of the manuscript.

Associate editor:
Sonia Regina Pasian

Received: Oct. 20, 2022

1st Revision: Jul. 23, 2023

2nd Revision: Aug. 19, 2023

Approved: Aug. 30, 2023

How to cite this article:
Cruz, A. C. A., Araújo, A. G. R., Hutz, C. S., & Bosa, C. A. (2023). Cognitive interview in the search for validity evidence based on response processes. *Paidéia (Ribeirão Preto), 33,* e3336. doi:https://doi.org/10.1590/1982-4327e3336