

Survey Item Validation

Melissa G. Wolf^{1*}, Elliott Ihm¹, Andrew Maul¹, and Ann Taves¹

¹University of California, Santa Barbara

*Corresponding author: melissagordon@ucsb.edu

This chapter has been accepted for publication and will appear in a revised form in the Handbook of Research Methods in the Study of Religion (2nd ed.). Please carefully note that subsequent versions of this manuscript may have different content.

Abstract

In the social sciences, **validity** refers to the adequacy of a survey (or other mode of assessment) for its intended purpose. **Validation** refers to the activities undertaken during and after the construction of the survey to evaluate and improve validity. **Item validation** refers here to procedures for evaluating and improving respondents' understanding of the questions and response options included in a survey. Verbal probing techniques such as cognitive interviews can be used to understand respondents' **response process**, that is, what they are thinking as they answer the survey items. Although cognitive interviews can provide evidence for the validity of survey items, they are time-consuming and thus rarely used in practice. The Response Process Evaluation (RPE) method is a newly-developed technique that utilizes open-ended meta-surveys to rapidly collect evidence of validity across a population of interest, make quick revisions to items, and immediately test these revisions across new samples of respondents. Like cognitive interviews, the RPE method focuses on how participants interpret the item and select a response. The chapter demonstrates the process of validating one survey item taken from the Inventory of Non-Ordinary Experiences (INOE).

1 Validating Surveys

Social scientists commonly use self-report surveys to collect information about individuals' beliefs, attitudes, and behaviors. When designed well and interpreted appropriately, surveys can be used to gather useful data from large samples at relatively low cost. However, constructing a high-quality survey is challenging even under the best of circumstances, and it requires rigorous quality-control procedures to ensure that responses to survey items can be interpreted and used as intended. These procedures are all the more critical when the topic of the survey is something as complex and culturally- and contextually-bound as religion. A great deal of care must be taken to ensure, in particular, that diverse participants understand the meaning of survey items in the way that researchers intend. Since most scholars in the humanities would assume that comprehension is crucial, methods for acquiring evidence of validity based on how participants interpret and respond to survey items offer a promising bridge between the humanities and the social sciences. This chapter explains the importance of validity within the overall validation process, presents newly developed methods for assessing it at the item level, and discusses the implications for survey research on religion.

2 Validation

In the social sciences, validity broadly refers to the adequacy of a survey (or other mode of assessment) for its intended purpose (for a quick overview of validity theory, see [7]). Validation refers to the activities undertaken during and after the construction of the survey to evaluate and improve validity. The contemporary literature on validity (e.g., in the *Standards for Educational and Psychological Testing*, [1], and by [5])

stresses (a) that validity is not an inherent feature of a survey (or other instrument), but rather a characteristic of the survey with respect to a particular use, and (b) that, as a consequence, validation is necessarily fit-for-purpose, such that different forms of argumentation and evidence may be necessary depending on the design and intended purposes of the survey.

Despite the potential for variation in light of design and purpose, validation of self-report surveys in practice follows a fairly standard model regardless of the survey's content or goals, focusing primarily on what the *Standards* refer to as validity evidence based on internal structure (e.g., Cronbach's alpha, factor analysis, item response theory) and relations to other variables (e.g., correlation, regression, structural equations; [9]). As specialists in validation and methodology point out, the standard practices test only a limited range of hypotheses relevant to validity (see, e.g., [2]). In particular they do not provide meaningful feedback about the respondents' thought processes when responding to survey items. Further, when offered as the sole evidence of validity, these methods can easily lead to 'false positives' in which even meaningless survey items may appear unproblematic according to traditional criteria ([6])¹. The standard practices also make assumptions about cross-cultural consistency that researchers in the humanities may find problematic. Thus, within the educational and cognitive psychological assessment communities, it is often thought desirable to create items that are universally understood and invariant across cultures, race, and gender. Conversely, within the humanities, systematic differences in item and construct interpretations across cultures are often expected and may be of primary interest to researchers. These issues, taken together, point to the need for methods that can provide evidence that survey items and response options are well- and comparably-understood by respondents, across the full range of cultural and linguistic settings in which the survey will be used, in a manner consistent with the expectations of the survey designers.

3 Validating Survey Items and Response Options

A substantial body of literature already exists on the best practices for construction of survey items (for an accessible summary, see, e.g., [4]), as well as on methods for acquiring what the *Standards* refer to as validity evidence based on response processes (see, e.g., [10]). In particular, cognitive interviews (or 'think-alouds'), in which respondents are asked to verbally describe their thought processes when reading survey items and deciding how to respond, can provide a wealth of valuable information about the validity of a survey ([3]). However, such methods are time and labor intensive and present additional challenges when conducted across cultures and without specific training in interviewing. Thus, researchers need more efficient and accessible methods for evaluating response processes. The balance of this chapter presents a technique that helps fill this need, and discusses the implications for survey research on religion.

3.1 The Response Process Evaluation Method

The Response Process Evaluation (RPE) method is a newly developed technique that can be used to help establish evidence of validity of self-report items. It turns cognitive interviews into open-ended meta-surveys that enable researchers to rapidly collect evidence of validity for each survey item from each population of interest, make quick revisions to items, and immediately test the validity of these revisions across new samples of respondents².

Like cognitive interviews, the RPE method focuses on how participants *interpret* the item and select a *response*, which taken together constitute the 'response process.' Both the *interpretation* and the *response* are necessary to understand how people process items. To assess an item's validity, we ask a series of meta-questions (MQs) about each survey item. The MQs, which mimic the prompts that would be used in a cognitive interview, are designed to elicit evidence that enables the researcher to evaluate if the item was understood as intended. For example, to determine if a person correctly interprets a survey item, we

¹Such methods also presuppose that the survey has been constructed for the purpose of measuring one or more quantitative psychological properties, sometimes referred to as 'constructs' or 'latent variables' (e.g., religiosity, altruism), and thus are not appropriate methods of validation for surveys that have other goals, such as when the primary focus is on interpretation of responses to individual items.

²If researchers intend to make cross-cultural comparisons, the survey can be translated and the RPE method can be used to validate the items on-line in more than one cultural context. Simultaneous validation allows researchers to revise items to ensure they are understood as intended without privileging one context over the other.

could ask: 'What do you think this item means?' To understand their response, we could ask: 'How would you respond, and why?'

Participants give open-ended responses to each MQ, which are then evaluated by the researchers, who may mark each participant's response to each survey item as 'understood', 'not understood', 'not enough information', or otherwise flagged as important for discussion. Researchers – who should be subject experts that know the intended interpretations of the items and uses of the instrument – evaluate the open-ended responses to each MQ in small batches (in our work, we have found that having roughly five participants provides adequate evidence) with at least two reviewers per batch. Reviewers then compare their evaluations. Survey items that are evaluated by both reviewers as 'understood' by all participants in the initial batch are then given to another batch of participants; items that are marked as 'not understood' or 'not enough information' are considered for revision and either re-administered in the next batch for more testing in their current form or revised and then re-administered. This goes on iteratively until the final version of the survey item has been evaluated an adequate number of times (in our work, we have found that twenty is optimal balance between sample size and cost). Researchers should substantiate revisions by documenting both the changes that led to improvements in item interpretation and the evidence that their final items were understood as intended.

If the intent is to make cross-cultural comparisons, the RPE method must be completed for each distinct population of interest. We believe that survey items should not be translated verbatim but instead written in such a way that they share the same meaning within each population (insofar as a shared meaning is possible). Results from the RPE method should be triangulated across cultures to arrive at a final survey item for each population that ideally is both culturally sensitive and demonstrates evidence of similar interpretations and response processes across groups. In practice, there are five possible outcomes of this validation process for each survey item: (1) one final item that has been commonly understood across both cultures and is cross culturally valid; (2) two culturally sensitive versions of the same item that are commonly understood across both cultures and are cross culturally valid; (3) one final item that is understood differently across cultures and is intraculturally but not cross culturally valid; (4) two final items that are understood within but not across cultures and are thus intra but not cross culturally valid; (5) the removal of the survey item entirely from the survey in one or both groups, if (e.g.) intra-cultural validity cannot be established. Establishing cross-cultural stability of a survey item (outcomes 1 or 2) enables researchers to confidently compare the same concept across cultures. Intra-cultural validity (outcomes 3 and 4) allows researchers to interpret responses to a survey item within a given context and to highlights meaningful differences in interpretation between contexts.

The RPE technique is an efficient alternative to verbal probing techniques, such as cognitive interviews and think-alouds. Like the verbal techniques, it orients researchers to their participants' frames of reference and allows them to refine the concept that they intend to assess based on the range of interpretations and response rationales persons give for each item ([8]: 57). In the illustration that follows, we see that the RPE method not only provides evidence of validity for a specific population of interest (English-speaking Americans) but also aids the researchers in clarifying concepts and revising survey items.

3.2 Utilizing the Response Process Evaluation Method

To illustrate the RPE technique, we will demonstrate the process of validating one survey item taken from the Inventory of Non-Ordinary Experiences (INOE). The INOE asks respondents whether they have had a series of experiences, e.g., 'I have had an experience in which it seemed as if there was another self in my body'. If a respondent indicates they have had an experience, they are asked follow-up questions regarding the context, significance, valence, and effect of the experience on their life. The experience items are expressed in generic terms that we hope will be understandable across cultures, and complex cultural concepts, such as religious or *dhaarmik*, are limited to the follow-up questions regarding how the experience was appraised. We use the RPE method here to demonstrate intra-cultural validation of one experience item in English-speaking Americans, but the method can easily be expanded to test cross-cultural validation of all survey items.

Many of the INOE's experience items were adapted from existing measures of religious, paranormal, and psychotic experiences. Others were added to capture experiences that are appraised as religious or spiritual in some contexts. The item we are using to illustrate the RPE process is of the latter sort. It was added to assess devotion to objects and figures, whether people appraised them as religious or not. In formulating

the item, we had in mind religious objects, such as relics, statues, and communion wafers, and figures, such as gurus and saints. The item initially read: 'I have had an experience of reverence or deep attachment toward an individual or object that stood out from all other such experiences' (Table 1).

Table 1: Initial INOE survey item and instructions.

Item Instructions	Please indicate whether or not you have had each kind of experience, by selecting 'Yes' or 'No'. Only select 'yes' if you can remember at least one specific experience that stands out.
Item	'I have had an experience of reverence or deep attachment toward an individual or object that stood out from all other such experiences'

To collect data and establish intra-cultural validity, we utilized an international online survey platform called Amazon Mechanical Turk (MTurk; Keith, Tay, & Harms 2017) and limited the availability of the RPE meta-survey to MTurk participants to English-speaking Americans who reside in the United States. We left all other categories (such as race, ethnicity, and income) unrestricted, and collected this demographic information from each participant at the end of the survey. Participants responded to MQs about each item iteration in small 'batches' of roughly five participants. Participant responses were divided into batches to allow us to gather validity evidence iteratively and make changes to items without investing substantial resources on a version of an item that was not interpreted as intended.

We divided our MQs into the two main segments, here termed *interpretation* and *response*, with two supplementary questions (Table 2). The *interpretation* MQ asks the participant to paraphrase the survey item in their own words. We found that respondents did not always paraphrase both the content of the experience and the added specification that it 'stood out from other such experiences' in response to a single interpretation MQ. To get at both, we divided the interpretation MQ into two, asking first about the specific type of experience and then about the item as a whole (including the specification that it 'stood out'). The format of the two-part *response* MQ is straightforward, asking first how they would respond and then what response option they would select. From this, we get the participant's open-ended response (and thus, a reflection of how they relate to the survey item content), which we can then compare with the response option they selected. Together, the *interpretation* and *response* MQs allow the researcher to evaluate the participant's understanding of the survey item.

Of the two supplementary MQs, one is universal, and the other is more specific. The *feedback* MQ serves as a catchall for lingering thoughts from the respondents. The *example* MQ was specifically tailored for the INOE, but it may prove valuable for other surveys depending upon their intended uses. Because the INOE separates experiences and appraisals and is intended for use across cultures, the *example* MQ indicates what specific experiences they had in mind when paraphrasing a generically worded item (i.e., the item stripped of appraisals). We plan to use the examples to map the range of meanings associated with the generic item in different cultural contexts.

Table 2: Initial meta-questions, in order.

<i>Interpretation</i>
What does 'reverence or deep attachment toward an individual or object' mean to you? (MQ1)
In your own words, what do you think this entire item means? (MQ2)
<i>Example</i>
Please give an example of such an experience whether or not you've had one. (MQ3)
<i>Response</i>
How would you respond? (MQ4)
If these were the response options, which would you select? (MQ5)
<i>Feedback</i>
Is there anything you don't understand or would change? If so, what? (MQ6)

In Table 3, we present a sample of four responses from the first round of MQs. We evaluated the responses of Participants 1 and 2 as 'understood' because the participants demonstrated a clear grasp of the meaning of 'reverence or deep attachment' in both their paraphrase (MQ1 and MQ2) and their response process (MQ4). We agreed in assessing the responses of Participants 3 and 4 as 'not understood' and 'not enough

information', respectively, and flagged both; albeit for different reasons. We flagged Participant 3's paraphrase of the item in terms of 'awe' which, while plausible, did not quite fit with what we had in mind, thus causing us to reflect more precisely on what we meant by 'reverence or deep attachment.' Participant 4's response was marked as 'not enough information' because we did not feel that 'respect' was an adequate paraphrase of MQ1 and MQ2, even though the example of an heirloom (MQ3) was exactly the sort of object we had in mind. It was also flagged because they misinterpreted the prompt for MQ4, leading us to believe that it might need modification. The first round of MQs for the first iteration of the survey item were administered to 8 people; 75% of the responses were evaluated as 'understood'.

For the sake of space, we present just two more of the six iterations of this survey item: #3 and #6. In iteration #3, we decided to switch 'person or object' to 'object or person' because we noted that until that point, nearly all of the responses were about individuals rather than objects and we wanted to make sure people noticed the word 'object'. We added the word 'devotion' to see if that would help clarify the meaning of 'reverence', which appeared to be confusing some respondents. Thus, iteration #3 read: 'I have had an experience of *devotion*, reverence or deep attachment toward an *object or individual* that stood out from all other such experiences' (modifications in italics).

Table 4 presents a sample from the first batch of responses to iteration #3. Of these responses, two were evaluated as 'understood' and two were not. Participants 1 and 2 did not provide coherent responses to the MQs and subsequently indicated that they had not had such an experience (likely because they did not understand the survey item). Participants 3 and 4 were evaluated as having understood the survey item because their paraphrases were in line with what we intended, and their response processes were coherent. We noted an increase in the number of participants that gave examples of objects instead of persons, suggesting that switching the order may have been successful. We also noted that some respondents were giving general examples, such as 'completing my schoolwork' when we wanted people to think of specific objects or persons. The second batch of iteration 3 was comprised of 7 people; 72% of the responses were evaluated as 'understood'.

In light of responses to iterations 3-5, we added the word 'particular', dropped 'reverence' (because it appeared to confuse one in five participants), changed the item stem from 'experience' to 'feeling', and added emphases. Thus, as of iteration #6, our item read: 'I have felt devotion or deep attachment toward **one** particular object or individual that stood out from all other such feelings' (emphases are part of the item). Additionally, we moved the example MQ to the end to create a more natural flow for the respondents (making it MQ5 instead of MQ3) and rephrased MQ3.

The evaluation process for iteration 6, which turned out to be the last, began like its predecessors: with a batch of roughly five participants. When the first five were evaluated as 'understood' (100% understood), we ran another batch of five, and repeated the process until we had twenty responses to the MQs. In all, 19 out of 20 respondents correctly appraised the survey item (95% understood), with one evaluated as 'not enough information'; a marked increase in comprehension from earlier iterations. Four of the responses are presented in Table 5. All four respondents were evaluated as having understood the survey item because they seemed to have no trouble paraphrasing or responding to the meta-questions, and the examples they gave were logical and meaningful. If we had simply administered the initial survey item without validating it, about 25% of our respondents would likely not have understood the items as intended, making conclusions based on analysis of their data inaccurate. Also negative responses would have been overrepresented, since respondents tended to say they did not have an experience when they did not understand the survey item (see Tables 3 and 4).

Table 3: **Iteration #1:** sample responses from the first batch of respondents.

'I have had an experience of reverence or deep attachment toward an individual or object that stood out from all other such experiences' (75% understood).

Participant	MQ1: What does 'reverence or deep attachment toward an individual or object' mean to you?	MQ2: In your own words, what do you think this entire item means?	MQ3: Please give an example of such an experience whether or not you've had one.	MQ4: How would you respond?	MQ5: If these were the response options, which would you select?	Evaluation
1	an obsession or strong like or unhealthy concern with a certain person or thing	If you have ever had a strong, obsessive-like attachment to a person or an object before	had a crush on a boy throughout high school and stalked his house every day, devoted journals to thoughts about him, was secretly in love with him	no, i am positive i never have. i don't really get that attached.	No	Understood
2	deep love of a person or object.	There is someone I respect and love and care deeply about.	I believe I feel this towards my girlfriend. I really appreciate her, and will probably ask her to marry me one day. I love her.	I would say yes, there is someone I care deeply about.	Yes	Understood
3	I felt awe towards someone or something.	A moment of extreme awe	[left blank]	I can't think of anything.	No	Not understood; Flagged
4	It means that you really respect a person or object.	It means that you have felt emotions very strongly of respect towards a person or object.	People who perhaps receive an heirloom that is very important.	I would be very respectful and very serious because it is such an important thing.	No	Not enough information; Flagged

Table 4: **Iteration #3**: sample responses from the first batch of respondents.

'I have had an experience of devotion, reverence or deep attachment toward an object or individual that stood out from all other such experiences' (72% understood).

Participant	MQ1: What does 'devotion, reverence or deep attachment toward an object or individual' mean to you?	MQ2: In your own words, what do you think this entire item means?	MQ3: Please give an example of such an experience whether or not you've had one.	MQ4: How would you respond?	MQ5: If these were the response options, which would you select?	Evaluation
1	Never happen to me.	Never happen to me.	Never happen to me.	Never happen to me.	No	Not understood
2	obey something	It is very confusing	I have not had a similar experience	no	No	Not understood
3	It seems self-explanatory. Perhaps a person feels indebted to an individual or that they owe them something.	The item is simply asking for an experience from the participant.	Maybe someone had saved someone's life, and the person rescued feels devoted to their savior.	I would respond with no as I have not had such an experience.	No	Understood
4	It means if we have particular feelings for something or someone	If there is something or someone that we feel very near to us because of a sentimental reason	A son that inherits his dad watch?!	I'm very attached to a coin my grandpa gave me, I wouldn't give it away for anything in the world	Yes	Understood

Table 5: **Iteration #6:** a sample of responses.

'I have felt devotion or deep attachment toward **one** particular object or individual that stood out from all other such feelings' (95% understood).

Participant	MQ1: What do you think 'devotion or deep attachment toward one particular object or individual' means?	MQ2: In your own words, what do you think this entire item means?	MQ3: Do you think you've felt devotion or deep attachment toward one particular object or individual that stood out from all other such feelings? Why or why not?	MQ4: If these were the response options, which would you select?	MQ5: Please give an example of such a feeling whether or not you've had one.	Evaluation
1	Something you feel that you cannot go without. it is central to your existence.	Something that is attached to you almost as if it is a body part or something of that sort of importance.	I have as it is my mom's ashes and it stands out because it is all that I have left of her. Her urn and ashes are now objects to me that I cannot go without having near me and I feel are central to my overall well being and existence.	Yes	I have had one as when I am not near my mom's ashes and urn I feel like something is gone and it is not a part of me whatsoever. it makes me feel as if something is wrong with me and I hate the feeling as a whole.	Understood
2	I means you hold a specific object or person with a great personal significance.	Do you have an item or person that means more than anyone else ever?	No. I just haven't.	No	A parent to their first child.	Understood
3	it means I feel close or emotionally bonded to a person or a thing, in a powerful & unique way	I'm guessing it means that this feeling of closeness or attachment is so strong that I don't have that feeling for anything/anyone else. But it's worded strangely.	I feel that about Boracay, my favorite place in the world. In my eyes no other place in the world comes close to how much I love and enjoy it.	Yes	About Boracay. I love the place like no other. I still have that feeling about it.	Understood
4	Devotion to a person, is what i feel toward my 2 children. I am beyond awe that me and partner created these children, and totally committed to enabling them to be the best people possible	this is a unique and spectacular commitment of adoration to impart whatever is necessary with emotion and passion	the actual birth of my children produced an inherent devotion to other human beings	Yes	the feeling was the birthing experience	Understood

4 Conclusion

Using the RPE method, we were able to refine the INOE survey item, quantify and qualify the improvement in its interpretability, and find evidentiary support that it was understood as we intended within this specific population. Because the experience items in this survey are worded generically, it is possible that many can be worded in a way that is universally understood across cultures. This will be determined by repeating the validation process in other contexts and triangulating the RPE results. Items designed to assess these appraisals will be translated and validated using the same RPE process, but we expect that some of translated terms in the appraisals (e.g., religion and *dhaarmik*) will be interpreted in culture-specific ways. If experience and/or appraisal items are understood in the same way within a culture, but not across cultures, we will use the culture-specific understanding arrived at through the validation process to interpret our results. The RPE method not only provides evidence of validity for survey items within one population but enables humanists to evaluate the cross-cultural stability of complex constructs.

References

- [1] American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. 2014.
- [2] Denny Borsboom. "The attack of the psychometricians". In: *Psychometrika* 71.451 (2006), pp. 425–440. ISSN: 0033-3123. DOI: [10.1007/s11336-006-1447-6](https://doi.org/10.1007/s11336-006-1447-6).
- [3] Miguel Castillo-Diaz and Jose-Luis Padilla. "How cognitive interviewing can provide validity evidence of the response processes to scale items". In: *Social Indicators Research* 114.3 (2013), pp. 963–975.
- [4] Hunter Gehlbach and Maureen E. Brinkworth. "Measure twice, cut down on error: A process for enhancing the validity of survey scales". In: *Review of General Psychology* 15.4 (2011), pp. 380–387.
- [5] Michael Kane. "Validation". In: *Educational Measurement*. Ed. by R. L. Brennan. 4th ed. Westport, CT: ACE/Praeger, 2006, pp. 17–64.
- [6] Andrew Maul. "Rethinking Traditional Methods of Survey Validation". In: *Measurement: Interdisciplinary Research and Perspectives* (2017).
- [7] Andrew Maul. "Validity". In: *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*. Ed. by Bruce B. Frey. Thousand Oaks: SAGE Publications, Inc., 2018, pp. 1771–1775. ISBN: 9781506326153. DOI: [10.4135/9781506326139](https://doi.org/10.4135/9781506326139). URL: <http://methods.sagepub.com/reference/the-sage-encyclopedia-of-educational-research-measurement-and-evaluation>.
- [8] Mark Wilson. *Constructing Measures*. New York, NY: Taylor & Francis Group, 2005.
- [9] Bruno D Zumbo and Eric K. H. Chan. *Validity and Validation in Social, Behavioral and Health Sciences*. 2014, pp. 1–329. ISBN: 9783319077932. DOI: [10.1007/978-3-319-07794-9](https://doi.org/10.1007/978-3-319-07794-9).
- [10] Bruno D Zumbo and Anita M Hubley. *Understanding and Investigating Response Processes in Validation Research*. Vol. 69. Cham, Switzerland: Springer International Publishing, 2017. ISBN: 978-3-319-56128-8. DOI: [10.1007/978-3-319-56129-5](https://doi.org/10.1007/978-3-319-56129-5). URL: <http://link.springer.com/10.1007/978-3-319-56129-5>.