

Updates on a Question Appraisal Tool: The Case against Indirect Rating Tasks

Jack Fowler and Carol Cosenza

Center for Survey Research, University of Massachusetts, Boston

1. Introduction

For some time, we have been working on approaches to identifying question problems before cognitive testing so that obvious problems could be either flagged for evaluation during testing or, better yet, addressed before testing. A version of a question appraisal form was presented at QUEST 2003.

In designing this tool, one of our goals was to have standards that did not depend upon the opinion of the rater. So, we avoided guidelines such as “questions should be clear”, because clarity could not be unambiguously evaluated; it is somewhat in the eye of the rater. We wanted to share two advances in our conceptualization of question problems that we think may expand its value.

It has become clear to us that one primary source of ambiguous questions is the use of abstract nouns that refer to a class or events or activities. What is included in ‘exercise’, ‘news media’, ‘health providers’, ‘problems’? Although there certainly are other unclear words and concepts in questions, we think complex abstract nouns that are not defined are an identifiable problem that can be reliably identified.

Our previous version also identified agree-disagree and true-false questions as potentially problematic. We have developed a more generalized form of that question standard: questions designed to assign a rating or place something on a scale that adds a further irrelevant task to the question in order to evoke an answer. Agree-disagree is just one of a number of ways in which question designers do this. We think all such approaches, which we call “indirect ratings”, are potentially problematic.

2. The Appraisal Tool

The following is the current version of our Appraisal Tool. The goal of creating it was to identify a set of question characteristics that that could be reliably coded and that have demonstrable adverse effects on ease of administration or data quality. **The material in bold type constitutes the proposed changes.**

SYSTEMATIC INSTRUMENT APPRAISAL

Comprehension Issues

C1) Does the question have a reference period (time)? This applies to any question for which the answer could reasonably be expected to vary from day to day, week to week, or month to month.

C2) Is the question hypothetical?

C3) Are there multiple questions being asked in a single question? (Is the question multi-barreled?)

C4) Does the question include an abstract noun that is not defined?

Retrieval of Information

R1) Is the question cognitively complex? Does the question require multiple calculations in order to answer the question?

R2) Does the question contain assumptions about the respondent's situation, or the way the respondent thinks about things, that are not necessarily true but that are critical to answering the question?

Formation of Answer

F1) Does the question make the response task clear to the respondent; that is, is it clear what kind of answer is required, and at what level of detail, in order to meet the question objectives?

F2) (If this is a fixed-response question) Are the answer categories mutually exclusive and exhaustive?

F3) DELETE: Is the question an “agree-disagree” question, or a variation thereon (such as true-not true)?

REPLACE WITH: Does the question give respondents a task other than a direct rating to provide information about where something (an idea, experience, person, or institution) is seen to lie on some continuum?

Usability Concerns

U1) Is the question fully scripted, including when and how to use any optional text?

U2) Does the question end with a question? (Are definitions and introductory phrases at the beginning of the question?)

U3) Are there appropriate “skip” instructions so that respondents are asked to answer only those questions that apply to them?

U4) Are the response tasks that respondents are supposed to use appropriate to the question that is asked?

The notion of flagging abstract nouns, nouns that designate a variety of more specific things, events or characteristics, seems fairly straightforward. We plan to do some experimentation to see how reliably coders can agree on when such nouns are in questions.

The concept of an indirect rating is probably less familiar, so we will elaborate on that in more detail.

3. More about indirect ratings

Many survey questions are designed to have respondents place their perceptions or evaluations on some kind of continuum. The typical task is to ask respondents to choose the number or the adjective on the scale that best describes their assessment.

While evaluations based on a continuum from good to bad may be the most common, ratings can be made of promptness, ability, energy levels, or political conservatism. In each of these cases and many more, a continuum can be defined and respondents can be asked where on that continuum they think something lies. We will call questions like that “direct ratings”.

Direct Example: *In general, would you say your views on national issues are very conservative, somewhat conservative, mixed, neither liberal nor conservative, somewhat liberal, or very liberal?*

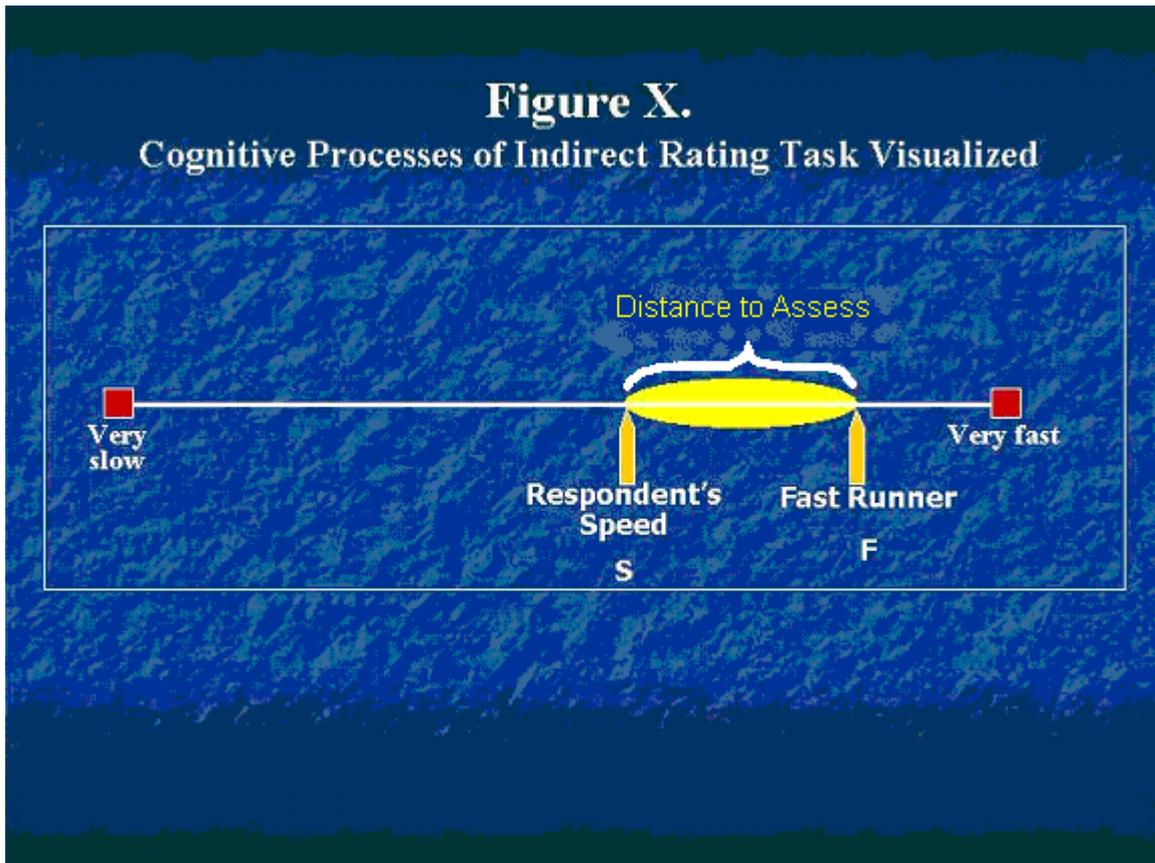
There is another approach to measurement that seems to accomplish the same thing. We call this an indirect approach to rating. The defining characteristic of such questions is that the stem of the question itself defines a spot on a continuum. Respondents are then asked how close that spot is to the way they see things.

Indirect example: *Do you consider your self to be a fast runner?*

In order to answer this question, a respondent must go through the following steps:

1. Decide how fast one must be able to run in order to be considered **fast**. Essentially, locate **fast** on a scale from very slow to very fast. This is designated as F on Figure X.
2. Assess where on that same scale the respondent would put his own running **speed**—designated S on Figure X.
3. Calculate the distance from F to S.
4. Decide whether or not that distance is small enough to qualify for a “yes” answer.

Figure X provides a picture of what a respondent has to go through in order to answer a question like that.



Let's take another example, which conceptually is exactly the same. In this case, compare the indirect question with the direct question about rating political views that was given above.

Indirect example: *Politically, do you consider yourself to have conservative views on national issues?*

This question, like the question about the fast runner, can be answered with a “yes” or “no”. The task for the respondent cognitively also mirrors the task for the fast runner question. First, the respondent has to decide where on the continuum from very liberal to very conservative to rate his views. Then he has to calculate how close his rating is to “conservative”, the point on the continuum specified by the stem of the question. Then, he has to evaluate the distance between the rating he would give his own views and “conservative” and decide if they are close enough that he is willing to say they are the same (or close enough to the same to produce a “yes” answer.)

Researchers have taken the complexity even further by designing a variety of ways to scale how close the statement in the question is to the respondent's perception of the true answer. The most straightforward approach is to give those who are feeling uncomfortable with an unqualified “yes” response the option of giving a qualified answer:

Example a: *Yes, Yes, to some extent, or No*

With a little bit of rewording, a four-category response scale can be offered to respondents, such as:

Example b: *Completely true, mostly true, somewhat untrue and completely untrue*

Finally, perhaps the most widely used response task is some variation on asking respondents whether they agree or disagree with some statement, using something like the following:

Example c: *Strongly agree, agree, neither agree nor disagree, disagree, strongly disagree*

We would assert that any and all of these indirect approaches to getting respondents to locate their views or feelings on a rating scale are less satisfactory than asking them to do it in a more straightforward way. So compare the yes/no and direct rating versions of the political question:

1. Indirect example: *Politically, do you consider yourself to have conservative views on national issues?*

2. Direct Example: *In general, would you say your views on national issues are very conservative, somewhat conservative, mixed, neither liberal nor conservative, somewhat liberal, or very liberal?*

Even though the first indirect alternative (yes/no) is simplified by having only two response options, it suffers from the cognitive complexity outlined above. In addition to being easier for respondents to answer, the direct rating also provides much more information, sorting respondents into 5 categories, rather than only two. It eliminates an important source of error variance between respondents in how close to “conservative” their perception of the true answer has to be before they will say “yes”. Example b above does not address any of these issues.

3. Indirect Example: *How true would you say it is that your views on national issues are conservative—would you say completely true, somewhat true, somewhat untrue, completely untrue.*

This version has all the problems noted in the examples above, plus two other issues. First, the scaling task becomes more problematic. What is the difference between “somewhat true” and “somewhat untrue”? It is easy to argue they are conceptually indistinguishable. Thus, it is also easy to think that people giving either of these answers may be very similar, differing mainly in the way they use the response alternatives. Second, consider the person who thinks he is “extremely conservative”. In order to analyze the answers to example 3, it is necessary to assume that those who give “less true” responses are less conservative. However, what if a respondent thinks that “conservative” understates his degree of conservatism? If that respondent said “somewhat true” to indicate an imperfect match, those in that category would include those who were both more and less conservative than conservative, and there is no longer a clear order in the answers. So, a further limitation of this indirect approach is that it is essential to place the point specified in the stem of the question at some extreme point on the continuum, so that demurring responses can be unambiguously interpreted as being in a particular direction.

4. Indirect Example: *Consider the statement that your political views on national issues are conservative. Would you agree strongly, agree, disagree, or disagree strongly?*

All of the issues raised above apply to questions in the agree-disagree form. In addition, there are two other concerns. First, it is difficult to pose agree-disagree alternatives that constitute a clear monotonic continuum. The form of this example includes the concept of the strength of feeling about the agreement, as well as simply how close the statement is to the respondent's view. It is not clear that the "strongly" responses mean they are closer than simply "agree" to the respondent's view, and it is common to analyze the results as a dichotomy: agree vs. disagree. There are other forms of the response categories (for example, completely agree, somewhat agree, somewhat disagree, completely disagree), but they all require that the specified spot on the continuum that is included in the question stem be at some extreme on the continuum in question, so that demurring responses can be unambiguously interpreted as being in a particular direction.

A further concern is acquiescence. It has been shown that some respondents are more likely to agree than disagree. Those who are less educated or have less knowledge of the topic are particularly likely to show this pattern. Acquiescence thus becomes another source of error variance, something that affects answers that has nothing to do with the true answer to the question.

Thus, the built in cognitive complexity, the difficulty of creating meaningful monotonic scaling categories, which in turn limits the number of categories that can be used in analysis, and the introduction of acquiescence bias all should lead researchers to avoid indirect rating tasks and prefer direct rating tasks when designing questions.

4. Empirical results

Over the years, we have found a number of problems with these indirect measures in various kinds of testing we have done—both cognitive testing and behavior coding of pretests. We will give three examples here.

One series of questions asked how likely people would be to sign up for recycling at various levels of cost to them. The idea was to use the amount they would be willing to pay as a measure of their commitment to recycling. However, cognitive testing revealed various problems with the underlying assumptions. For example, those who did not pay for trash collection seemed confused by the entire framework of the question. So did those who thought the city should pay them, or reduce their trash collection costs, not because they were not interested in recycling, but just because a city that is paid for recycled material should not be charging its residents. Thus, the assumption of a direct relationship between willingness to pay for recycling and commitment to recycling proved to be ill-founded and a source of error in the answers. They should have just asked people to directly rate how committed they were to recycling.

In a similar vein, teachers were asked how hard they wanted their union to work on a series of issues. The assumption behind the question form was that respondents would consider the importance or need for change in each issue, then consider how likely it was that the union could affect a positive result, then arrive at their answer. When we did cognitive testing, however, we found that many respondents could not deal with the two-step process the question design implied and required. Most simply ignored the issue of what was an appropriate role for union action and just rated their own sense of the priority of the issue. The questions would have been simpler if they had just asked for that kind of rating directly.

Finally, we compared an agree-disagree form of a series of three questions to a direct rating version of the same questions.

Version A. *The more medical tests people get, the healthier they are. Do you strongly agree, agree, probably agree, probably disagree, disagree, strongly disagree?*

Version B: *How big a role do you think medical tests play in keeping people healthy-- big role, a small role, or no role at all?*

We then behavior coded pretest interviews, half of which used each version. With respect to need for probing and repeating questions, the agree-disagree form of the question required probing 41% of the time versus 27% for the alternatives. When the items were correlated with a summary question to evaluate construct validity, the correlation of the direct ratings were all higher than those of the agree-disagree form, .27 on average versus .05.

5. Conclusion

The results presented above are typical of the kind of results we routinely get when we look at indirect approaches to measuring things that could be directly rated. One way to think of the problem is psychometrically. For a direct rating, classic theory says the answer reflects the true score and some error term:

$$x = t + e_d$$

Where x = the response

t = the true score

e_d = error associated with the direct rating task

When the response task requires the respondent to also calculate something else in addition to the basic rating, reflecting the indirect task, the equation becomes:

$$x = t + e_d + e_i$$

Where e_i = error associated with the additional indirect task rating.

Unless the indirect task is performed with complete consistency by all respondents, there will be more error in the measurement with an indirect task than with a direct rating—and our testing consistently shows that those indirect tasks tend to be particularly confusing and, hence, error prone.

So, based on these analyses, we believe we are justified in broadening the criteria in our appraisal form to try to flag all questions that use an indirect task. We are convinced that researchers would almost always be better served by using a more direct rating form of a question.