Questionnaire science

DL Oberski

Tilburg University, The Netherlands

Abstract

Some textbooks on questionnaire design accuse it of being an art. That would make the criterion for a "good" question entirely subjective–a worrying conclusion given that surveys are often used to discover important facts about people. Are our discoveries about people then also entirely subjective? This chapter shows that it is possible to study what a "good" or a "bad" question is by experimentation. Not only that: there is already a body of scientific evidence on questionnaire design that can, and should, be taken into account when designing a questionnaire. I review some of this evidence and show how it can be used to the advantage of the survey researcher. Of course, questionnaire science is far from complete. On the one hand, this means that some of our conclusions may still be more art than science. On the other, it means that we can agree on one aspect of questionnaire science: more of it is needed.

1 Why it is important to ask good questions

In polling, everything hinges on asking good questions. If I tried to measure your opinion about the current president by asking "how much do you like ice cream?", I would not get very far: that question would have no validity. But even if I did ask your opinion about the president, but did so in such a convoluted way that you would not know what to make of it, your answer might not be as valuable as it could have been. Take this made-up question, for instance:

> To which extent do you disagree with the statement "the current president's actions are not entirely unlike my own actions sometimes but some of his policies are not often bad."?
>
> 2 Not entirely disagree
>
> 3 Disagree
>
> -1 Don't know
>
> -2 Agree somewhat
>
> -3 Agree slightly

-4 Neither agree nor disagree

Is the statement about the president positive or negative, and to what extent? What "actions" and "policies" come to mind? Which is stronger: "somewhat" or "slightly""? Is category -1 neutral? Just a few of the many issues plaguing this unfortunate survey question. When you answer the question, you need to solve these issues to get to answer, but since the solutions are ambiguous at best, different people will choose different answer strategies – even if they had the same opinion about the president. If you changed your mind about the president next year, you might even solve the problem of answering this terrible question differently and give the same answer as you did previously, even though you changed your opinion. Such differences in answers between people with the same opinion are called "unreliability" in the literature (Lord and Novick, 1968). So even when a question is about the right topic, the way it is asked still determines how reliable the answers will be.

Unreliability is important because it strongly biases estimates of relationships (Fuller, 1987; Carroll et al., 2006). For example, if I were interested in the relationship between presidential approval and consumer confidence, I might calculate a correlation between these two variables; unreliability would then attenuate this correlation downwards, while common method variance would spuriously increase it. So this estimate would be severely biased and without additional information about the reliability and common method variance, there is no way of knowing the size and direction of this bias.

Unreliability's effects on estimates of relationships extends to relationships over time, such as panel or longitudinal data and time series (Hagenaars, 1990). Random measurement error will cause spurious shifts in opinion and jumps in time series that are purely due to the measurement error. Common method variance, on the other hand, can make opinions appear much more stable than they truly are.

When comparing groups, the measurement error resulting from poor question design may again bias the analysis. For instance, prior research suggests that highly educated respondents tend to "acquiesce" less – to agree to a statement regardless of its content (Narayan and Krosnick, 1996).

If we then compared the average response to an agree-disagree question in Washington DC, where 49% of adults hold a bachelor's degree, to West Virginia, where only 17% do[1], on average we would expect the West Virginians to agree more with any statement, regardless of its content. A researcher who found that Virginians indeed agreed more with her statement would then be at a loss to say whether this was because of a difference in opinion or one of measurement error. This incomparability is also called "measurement non-invariance", "measurement non-equivalence", or "differential item functioning" in the literature (see Oberski, 2012).

My contrived example serves to illustrate how unreliability may result from a question's phrasing and other characteristics, and that this unreliability is vital to draw accurate conclusions about many social phenomena. Of course I purposefully broke every rule in the book when phrasing the above question. Real polling questions follow "best practices", a set of approximate rules handed down by textbooks, or they are designed by experts. Even so, differences in respondents' answering strategy still occur, with the resulting unreliability of answers. And how can we be sure that all the many issues that could plague a survey question are actually taken care of in its formulation? Is expert opinion enough?

The remainer of this chapter aims to answer these questions. I argue that deferring to textbooks and experts is not enough to design the best questions, but that a body of scientific knowledge about questionnaire design does exist, comprising cognitive theory, empirical observations, and carefully designed experiments. I then discuss some examples of scientific knowledge about questionnaire design, including a large meta-analysis that has yielded user-friendly software encoding such knowledge.

2 What we do not know about asking questions

Pollsters and other survey research agencies have vast amounts of experience doing surveys. Thanks to these researchers' awareness that everything hinges on asking good questions, it has

[1] http://en.wikipedia.org/wiki/List_of_U.S._states_by_educational_attainment

Book	Negative	Categories	Agree-disagree	Double-barreled	
Bradburn et al. (2004)	Avoid (p. 325)	7 (p. 331)	Good (p. 244)	Bad	
Dijkstra and Smit (1999)	Avoid (p. 83)	-	Avoid (p. 95)	Bad	
Dillman (2011)	Avoid (p. 73)	-	Avoid (p. 62)	Bad	
Folz (1996)	-	-	Neutral	Bad	-:
Fink (2009)	Avoid (p. 29)	4 or 5	Neutral	Bad	
Fowler (2014)	-	-	Avoid (p. 105)	Bad	
*Marketing Scales**	50%	5, 6, or 7	67%	60%	

The aspect is mentioned, but no negative or positive advice is given.
* Based on a random sample of 10 scales from the book (s.e. about 15%).

Table 1: Best and actual practices for four commonly discussed question characteristics.

become common practice to vet the questions in advance using questionnaire reviews, pretests, and other such evaluations (see Madans et al., 2011, for an overview). These procedures are meant to ensure that the right questions are asked in the best way possible. Regardless of the procedure followed to improve a question, though, the initial design typically follows "best practices"– standards for designing survey questions that have become encoded in the many textbooks now available on good questionnaire construction.

So what practices are currently considered "best", and how many of them do survey researchers actually implement? To demonstrate this, I picked up a selection of well- and lesser-known "how-to" advice books on survey and questionnaire design, as well as the very comprehensive *Handbook of Marketing Scales* (Netemeyer et al., 2011), which contains over 150 meticulously documented examples of vetted questionnaires used in marketing research. Table 1 shows what these books advise regarding negative questions in a battery ("Negative"), the preferred number of categories ("Categories"), the use of agree-disagree questions ("Agree-disagree"), and double-barreled questions. These examples are by no means an exhaustive list of possible design choices, but are all commonly mentioned in the textbooks and serve to demonstrate how question design advice is given and taken.

Table 1 shows that, broadly, there is a consensus on some of these best practices, while others are contradictory. For example, all textbooks in the Table agree that double-barreled questions are a bad idea, and most agree that negatively formulated questions are to be avoided. On the other hand, there is little agreement between these authors on the use of agree-disagree questions or the number of categories: here, one author's best practice is another's *faux-pas*.

The bottom row of Table 1 is meant to give an idea of the actual–as (possibly) opposed to "best"–practice of marketing research surveys from a small sample of the scales in the *Handbook*. Where textbook authors agree on the "best" practice, the *actual* practice is more often than not the opposite: for example, I found double-barreled questions in 60% of the sampled scales and about half of the scales use the negative formulations that textbooks agree should be avoided. Moreover, there was very little actual variation in the number of scale points, most scales using seven-point scales: here there is a common practice even though a best practice is not actually agreed upon by the textbooks. A researcher following Bradburn et al.'s advice (p. 149) to take existing questionnaires as a starting point may then be forgiven for thinking that seven-point scales represent a consensus best practice.

While very limited, the microreview offered by Table 1 suggests that (1) some "best" practices are contradictory; (2) some consensus best practices are not usually followed; and (3) a strong common practice may be present, absent of any actual consensus on the best practice. In short, to quote Dillman (2011, p. 50) "the rules, admonitions, and principles for how to word questions, enumerated in various books and articles, present a mind-boggling array of generally good but often conflicting and confusing directions about how to do it"; deferring to common or "best" practices is clearly not enough to warrant trustworthy conclusions from our surveys.

3 Beyond agreeing to disagree: what we do know

If best practices are so conflicting, is question design a matter of taste? After all, the title of one of the most classic of all question design textbooks, Payne's *The art of asking questions* (1951), directly suggests exactly that. And if true, this arbitrary nature of survey question design would detract from the trustworthiness of conclusions based on such questions. Fortunately, though, we can decide which practices truly are "best" under the given circumstances by experimenting with them and there is now a substantial literature arbitrating between such practices.

As an example, consider one of the design choices of some apparent contention among textbooks: the agree-disagree scales that proved so popular in existing questionnaires. There are three good reasons to think that agree-disagree scales are, in fact, a bad idea.

First, there are theoretical reasons. Cognitive psychology suggests that agree-disagree scales place an unnecessary cognitive burden upon the respondent that causes respondents to ”satisfice”; that is, to take shortcuts when answering the questions. Révilla et al. (2013) compared the process needed to answer an agree-disagree question such as ”to what extent do you agree or disagree that immigration is bad for the economy?” with that needed to answer an ”item-specific” question such as ”how good or bad for the economy is immigration?”. The latter, a well-known model of cognitive survey response suggests, is answered in several stages: comprehension of the question, retrieval of relevant information, judgment of this information, and response (Tourangeau et al., 2000).

In the example question ”how good or bad for the economy is immigration?”, the respondent would first read and understand words such as ”immigration”, ”economy”, ”good”, and ”bad”, as well as the grammatical structure of the sentence which gives it meaning – for instance the presence of the WH word ”how”, turning the phrase into a request for graded information. If the respondent is satisficing, the phrase might not be read, but the answer categories might be read directly instead. These might say something like ”immigration is very good for the economy”, a sentence that communicates the required meaning on its own. Subsequently, information stored in memory about relevant concepts is retrieved until the respondent has had enough. When satisficing, the respondent may only retrieve the most salient information: things that they may have heard just recently or very often. In the next stage, the theory suggests, this information is weighed and the actual opinion formed. Again, instead of weighing all the pros and cons as a professional economist might do, a respondent trying to get through the questionnaire may use simple rules to reach their judgment. Finally, the opinion must be mapped onto the response scale. If the respondent's internal idea about their opinion matches the labels closely, this can be a matter of ”choosing the option that comes closest", as we often instruct our respondents. A satisficing respondent may choose a different strategy. For example, they may choose one side of the issue an opt for the most extreme response on that side. This is known in the literature as "extreme response style". Thus, at each stage there is a potential for satisficing.

Our hypothetical journey through a survey question-and-answer process shows that answering a question is a complicated cognitive process. Because it is so complicated, different respondents holding the same opinion could give different answers. The higher the cognitive burden of answering a question, the more respondents will satisfice, and the more their answers will differ erroneously and correlate spuriously.

And that is precisely the theoretical problem with the agree-disagree format such as "to what extent do you agree or disagree that immigration is bad for the economy?": its cognitive burden is higher than that of the direct question. At the response stage, it is not enough for the respondent to simply find the response option closest to her opinion. Instead, she must create a mental scale of opinions, locate the statement on it, locate her own opinion on it, and then decide how the distance between them maps onto an agreement scale (e.g. Trabasso et al., 1971). If this process sounds incredibly burdensome, you are right. To avoid this burden, respondents often satisfice. Thus, we think that agree-disagree questions imply a higher cognitive burden because respondents take much longer to answer an agree-disagree question than to answer the corresponding direct question; and because, when they do, we observe more satisficing behaviors.

The psychologist Rensis Likert (1903–1981), who is often said to have invented agree-disagree questions, was well aware of this potential problem. His solution to the problem was to authoritatively assume it away: "it is quite immaterial what the extremes of the attitude continuum are called. (...) it makes no difference whether the zero extreme is assigned to 'appreciation of' the church or 'depreciation of' the church" (Likert, 1932, p. 48). We now know this to be false. Experiments show that varying the extremeness of the statement or negating it with the word "not", which Likert thought would not make any difference, can in fact radically shift the answers people give (e.g. Schuman and Presser, 1981). Worse still, the effect seems to differ over respondents, causing random errors.

This brings us to the second set of reasons to discard agree-disagree scales: they are less valid and less reliable than direct questions. ”Unreliable” means there will be variations in the answers of people who we suspect have the exact same opinion. After all, if two people have the same opinion, the ideal, perfectly reliable, opinion poll would yield equal answers. Similarly, known differences should be reflected in the answers. For example, a question about the role of women in society should at least on average be related to gender. A invalid question, which does not measure the intended opinion, will fail such tests.

Unfortunately, a person’s ”true opinion” cannot be observed. We can, however, translate the two requirements of reliability and validity into numbers that can be estimated from observable data. There are various approaches to doing so, all of which involve taking not just one but several measures of the same phenomenon to make statements about reliability and/or validity. Commonly used approaches are the quasi-simplex model (Heise and Bohrnstedt, 1970; Wiley and Wiley, 1970; Alwin, 2007, 2011), in which each respondent is asked the same question in multiple waves of a panel, and the multitrait-multimethod (MTMM) approach (Campbell and Fiske, 1959; Andrews, 1984; Saris and Gallhofer, 2007b; Saris et al., 2012), in which a within-persons experiment is performed on the question format. Various studies performed in several countries suggest that both the reliability and the validity of questions estimated in this way in an agree-disagree format are lower than that in other formats (Krosnick and Fabrigrar, 2001; Saris et al., 2010).

The third and final reason to discard agree-disagree scales might form an explanation for the empirical finding that these scales are less valid and reliable: acquiescence. Acquiescence is the empirical finding that ”some respondents are inclined to agree with just about any assertion, regardless of its content” (Révilla et al., 2013). For example, Krosnick (2009) reported that 62–70% of respondents agree with the question ”do you agree or disagree with this statement?”. This question measures nothing, but people lean towards agreeing with it anyway. Other studies have found that a sizable group of people will agree both with a statement and its opposite (e.g. Selznick and Steinberg, 1969). Furthermore, pointless agreement is more common in low-education groups, younger, and tired respondents (e.g. Narayan and Krosnick, 1996). So the tendency to agree with anything varies over respondents. This does not only create random differences between people, but also spuriously correlates any questions that are asked in the agree-disagree format, since part of their shared variance will be shared acquiescence.

The agree-disagree format is an example of a common practice on which survey design textbooks do not agree–even though the theoretical and empirical evidence against it, of which this section has only scratched the surface, is impressive. Reviewing that body of evidence is not a trivial task, however. What's more, the agree-disagree format is just one of the many choices a researcher is faced with when asking a question; the number of categories, use of negative formulations, and double-barreled phrases were already mentioned. But there are many more: whether to balance the request, for example by asking "is immigration good or bad for the economy?", rather than just "bad for the economy", is another example, famously studied by Schuman and Presser (1981). Other choices are the complexity of the sentences used, the grammatical structure of the sentences, whether to give further information or definitions to the respondent, where to place the question in the questionnaire, the choice of answer scale, the choice of labels if response categories are used, and so on.

To get a feel for these choices, refer to Figure 1, and–without reading the footnote–try to spot the differences between the three versions. Some are obvious, such as the number of scale points. Others less so. For example, versions A and C are very similar but could in fact be considered to differ on at least six aspects that the literature has suggested may matter for their reliability and validity[2].

Clearly the number of choices made whenever we ask a respondent a question is considerable. Figure 2 shows a number of these choices for which the literature has suggested that they make a difference to the reliability and validity of the question (Saris and Gallhofer, 2007a). While knowing of their existence is useful, this knowledge does not immediately lead to better survey

[2] In terms of the coding scheme on the next page, these are: direct question (C) vs. other (A); use of a WH word ("how"); complexity of the request (A has more words and more syllables per word); interviewer instruction (C); labels are numbers (C) vs. boxes (A); presence of a "don't know" category. There may be more.

Version A. The next 3 questions are about your current job. Please choose one of the following to describe how varied your work is.

□ Not at all varied
□ A little varied
□ Quite varied
□ Very varied

Version B. Please indicate, on a scale of 0 to 10, how varied your work is, where 0 is not at all varied and 10 is very varied. Please tick the box that is closest to your opinion

Not at all varied										Very varied
0	1	2	3	4	5	6	7	8	9	10
□	□	□	□	□	□	□	□	□	□	□

Version C. Now for some questions about your current job.
Would you say your work is... [*Interviewer: READ OUT*]

1 ...not at all varied,
2 a little varied,
3 quite varied,
4 or, very varied?
8 (Don't know)

Figure 1: Three ways to ask a question, all tried in the European Social Survey (2002).

questions: it would be an insurmountable task for a researcher to go through the literature on each of these issues or do her own experiments for every single question asked. Moreover, as the example in Figure 1 illustrates, it may not be so easy to recognize every single relevant choice made. Without a tool to code these choices, we are at risk of focusing on issues that happen to be highly studied or that experts happen to have a strong opinion on, to the possible detriment of other choices that are less eye-catching but equally crucial to obtaining adequate measures of peoples' opinions. What we need to make informed evidence-based decisions is a structured summary of the literature on these issues: a meta-analysis on what makes a better or worse survey question.

4 A meta-analysis of survey experiments

One such meta-analysis is a multiyear project we performed in 2011 (Saris et al., 2012) on several thousand questions that were a part of the European Social Survey, as well as others part of a project executed in the United States and several European countries (these questions were also

Topic:

- Domain
- Concept
- Social desirability
- Centrality to respondent
- Fact vs. opinion
- Past/present/future

Wording:

- Direct question vs. other formulations
- Period or date
- WH word used
- Use of gradation
- Balance of the request
- Encouragement in question
- Emphasis on subjective opinion
- Other peoples' opinion given
- Stimulus or statement
- Absolute/comparative
- Knowledge or definitions
- Avg. words/sentence
- Avg. syllables/word
- No. subordinate clauses
- No. nouns
- No. abstract nouns
- Introduction used
- Avg. words/sentence, intro
- No. subordinate clauses, intro
- No. nouns, intro
- No. abstract nouns, intro
- Avg. syllables/word, intro

Administration:

- Computer assisted
- Interviewer present
- Oral/visual
- Showcard used
- Showcard horizontal/vertical
- Showcard pictures
- Showcard letters/numbers
- Showcard labels overlap
- Interviewer instruction
- Respondent instruction
- Position in the questionnaire
- Country
- Language

Response scale:

- Type of response scale
- Number of categories
- Labels full, partial, or no
- Labels full sentences
- Order of labels
- Numbers correspond to labels
- Unipolar/bipolar: theoretical
- Unipolar/bipolar: used
- Neutral category
- No. fixed reference points
- Don't know option

Figure 2: Some choices made when formulating a question and coded in SQP 2.0.

included in Andrews, 1984; Scherpenzeel, 1995; Saris and Gallhofer, 2007b). Other analyses can be found in Alwin and Krosnick (1991); Alwin (2007). In this project, we followed the following steps:

1. Estimate the reliability and common method variance (together: "quality") of a large number of questions;
2. Code characteristics of the questions that literature suggests relate to question quality;
3. Predict question quality from question characteristics (meta-analysis);
4. Create a freely available online web application that allows researchers to input their question and obtain its predicted quality; the "Survey Quality Predictor" (SQP).

The following subsections briefly explain each of these steps, focusing most attention on the practical tool for applied survey researchers, SQP.

4.1 Estimating question quality

There are several possible indicators of how good a question is. Two highly important indicators of quality are the reliability and common method variance. Both reliability and method variance can be expressed as numbers between 0 and 1 and can be interpreted as proportion of variance explained (R^2) of true variance (reliability) and method variance, respectively.

The *reliability* of a question is the correlation that answers to the question will have with the true values (or "true score"). For example, when asking about the number of doctors' visits, reliability is the correlation between the number of times the respondents claim to have visited the doctor on the one hand, and the actual number of times they visited the doctor on the other hand. When dealing with opinions, a true value is difficult to define, and, instead, a "true score" is defined as the hypothetical average answer that would be obtained if the same question were repeated and there were no memory (for more precise explanations of these concepts see Lord and Novick, 1968; Saris and Gallhofer, 2007a).

The *common method variance* of a question is the proportion of variance explained by random measurement effects, such as acquiescence, that the question has in common with other, similar questions. This shared measurement error variance causes spurious correlations among question answers. For example, if a question has a common method variance of 0.2, it can be expected to correlate 0.2 with a completely unrelated question asked in the same manner ("method"; Saris and Gallhofer, 2007a).

Campbell and Fiske (1959) suggested an experimental design to study both reliability and common method variance simultaneously: the "multitrait-multimethod" (MTMM) design. Procedures to estimate reliability and method variance of survey questions directly using structural equation models (SEM) were subsequently applied by Andrews (1984). Each such experiment crosses three survey questions to be studied ("traits") with three methods by which these questions can be asked ("methods"). By applying decomposition of variance using SEM, we can then disentangle what part of the survey questions' variance is due to the question, what part is due to how it was asked, and what part is not reproducible across repetitions (random error). A deeper explanation of MTMM experiments from a within-persons perspective can be found in Cernat and Oberski (2017).

Already in 1984, Frank Andrews (1935–1992) suggested to perform not just one, but several multitrait-multimethod experiments on survey question format, and summarized the results by comparing the quality of questions in different formats with each other. Over a period of several decades, this idea was subsequently expanded and improved upon by Saris and his colleagues (Saris and Andrews, 1991; Költringer, 1995; Scherpenzeel, 1995; Oberski et al., 2004; Saris and Gallhofer, 2007a,b; Saris et al., 2010, 2012; Révilla et al., 2013). They performed hundreds of MTMM experiments, obtaining estimates of the reliability and method variance of thousands of survey questions. These efforts led to a large database of 3483 questions – among them the "job variety" questions in Figure 1 – on which about 60 characteristics that are thought to affect question quality in the literature have been coded. Most of these characteristics are shown in Figure 2. Not all issues are included, such as the usage of double-barrelled requests or negative formulations. However, many issues found in the literature are addressed in this coding scheme (see Saris and Gallhofer, 2007b, for more information on the coding scheme and its development).

4.2 Coding question characteristics

The questions were coded by two experts as well as a group of trained coders at the Pompeu Fabra University, Spain. The codes for questions in languages unfamiliar to the experts were compared to those for the English versions of the questionnaires, and any differences were reconciled. The resulting database of questions with their codes was cleaned and merged with a database of estimates of the reliability and common method variance from multitrait-multimethod (MTMM) experiments. In these experiments, each respondent answered two different versions of the same question with about an hour of interview time in-between – for example, versions A and B from Figure 1. The same respondent also answers different questions in these same versions A and B – for instance on satisfaction with wages and health and safety. By combining the answers to different opinion questions asked in the same way with different methods of asking the same opinion,

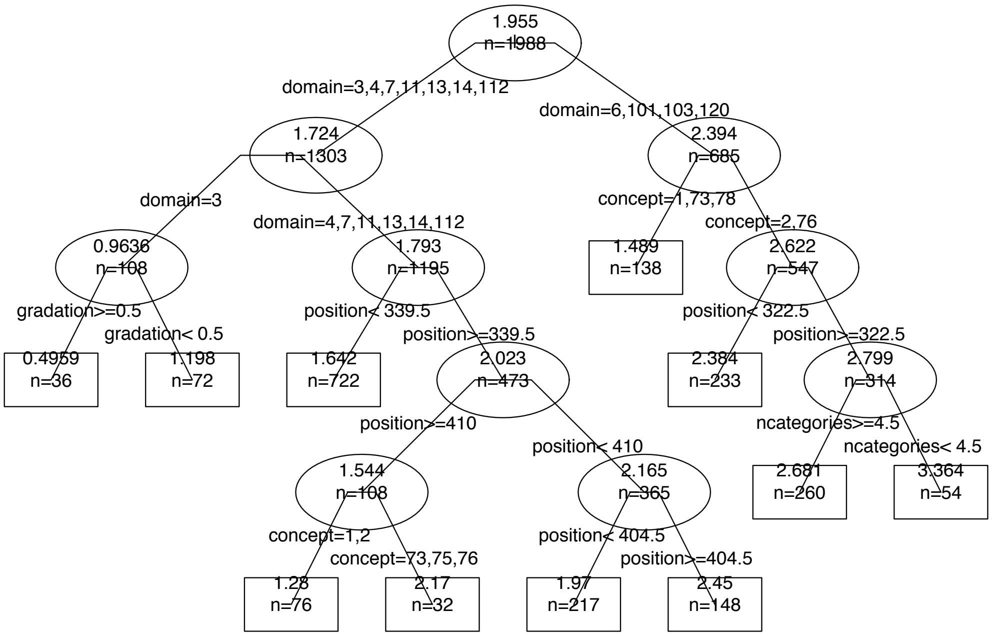


Figure 3: Example of a regression tree predicting the reliability of a question from a selection of its characteristics. The random forest consists of 1500 such trees.

confirmatory factor analysis can be used to separate the effects of the opinion (reliability) from those of the method (common method variance). Sometimes the complement of common method variance is called ”validity” in the MTMM literature. I will avoid that word here to avoid confusion with other, perhaps more familiar uses of that term. The end result was a large database of questions with two pieces of information: the MTMM reliability and common method variance, and the characteristics of these questions that might predict the reliability and method variance.

4.3 Predicting quality from characteristics

Machine learning techniques were then applied to predict the MTMM reliability and method variance of a question from its characteristics. By using random forests of regression trees (Breiman, 2001), 65% of the variance in reliability across the questions and 84% of the variance in the common method variance could be explained in questions that were in the ”testing sample”; that is, not used in the estimation of the model.

Figure 3 shows an example of one regression tree. The ”leaves” of this tree can be followed downwards, according to the characteristics of the question, to come to a prediction of the reliability (shown in logits). For example, the second-to-leftmost leaf in Figure 3 shows that a question on health issues (domain = 3) that uses a gradation in the question (”how much”, ”to which extent”), is predicted to have a reliability of invlogit$(1.198) = 0.768$, or about 80% reliability. There were 72 such questions in this training sample. These regression trees are, however, prone to overfitting. A random forest therefore randomly samples cases to be in either the training or testing sample. Furthermore, many variables may be strongly collinear (confounded) with one another. To counter this, the algorithm samples a random subset of the characteristics as well. This doubly random sampling is performed 1500 times, and a regression tree is learned on each of the training sets. Combining the 1500 predictions obtained from each of the trees by taking their average then yields the final prediction from the forest. The same procedure was applied to predict the common method variance.

The random forest yields a method that can predict the reliability and method variance of a question from its characteristics. However, following the procedure described here will be a tedious task for a survey researcher. This is why the results of the meta-analysis have been included in an online tool that is free to use. The following section describes this tool, developed to allow researchers to code their question characteristics and obtain a prediction from the random forest of the question’s reliability and common method variance.

4.4 Using the results of the meta-analysis to guide question design using the Survey Quality Predictor (SQP) 2.0

The Survey Quality Predictor (SQP) 2.0 (`http://sqp.upf.edu/`) is an online web application that is free to use. It has the following goals:

- Allow survey researchers to code their questions in the coding system of Saris and Gallhofer (2007a), becoming aware of the many choices made in designing a question;
- From the meta-analysis, predict the reliability and common method variance of the survey

question, so that the researcher can get an idea of the adequacy of their question for the research purpose.

- Tentatively suggest improvements based on the meta-analysis.

It does **not**:

- Estimate average *bias* in the question, for example due to social desirability;
- Predict other measures of a question's quality, such as the appropriateness of the question for the research topic or the number of missing responses;
- Include every possible characteristic of a question–although it does include many of them;
- Provide information about cause and effect; changing characteristics may not always result in the predicted improvement;
- Give highly accurate predictions for questions about behaviors and fact. The main focus has been questions on opinions, feelings, evaluations, and so on.

A final caveat is that SQP has not been tested extensively on questions in web surveys, although research suggests that web and other self-administration modes do not differ in reliability and method variance (Révilla, 2012a,b; Révilla and Saris, 2012), so that the predictions using self-administration as the mode may be reasonably adequate.

In spite of all these limitations, SQP can be a very useful tool for survey designers. To demonstrate the working of the program, I have coded Version A of the "job variety" question into the system.

The first step is to enter the question text itself into the system. Figure 7 (Appendix) shows that this text is split up into three parts: the introduction, "request for an answer", and answer scale. Each of these choices is explained on the page itself. As the name implies, the request for an answer refers to the request itself, while the introduction is any leading text such as "now for some questions about your health". After entering the question text, the coding system appears, as shown in Figure 8 (Appendix). Clicking the "Begin coding" button begins the coding process.

As Figure 4 demonstrates, the characteristic will appear on the left while coding, together with an explanation of it. The user then chooses a value, which is subsequently displayed on the right and can be amended at any time. Where possible, some characteristics are coded automatically. For questions asked in English and a few other languages, for instance, natural language processing (part-of-speech tagging) is applied automatically to the texts to count the number of nouns and syllables, as Figure 9 (Appendix) shows.

After finishing the coding process, some predictions are shown with their uncertainty. The reliability coefficient, “validity coefficient” (complement of the method effect), and their product, the so-called “quality coefficient” (Saris and Gallhofer, 2007a), are shown (Figure 5). The quality coefficient squared indicates the proportion of variance in the answers to the questions that we can expect to be due to the person’s true opinion. The reliability coefficient of 0.8 in Figure 5 suggests that any true correlations the answers to this question might have with other variables will be attenuated (multiplied) by 0.8. This includes relationships over time, so that any time series of this variable will jitter up and down randomly by at least 20% more than is the reality. A “validity coefficient” of 0.99 indicates that two questions asked in this same manner can be expected to correlate spuriously by a very small amount (this spurious additional correlation can be calculated from the “validity” coefficient as $1 - 0.985^2 = 0.0298$). Common method variance is therefore predicted not to be a great concern with this question.

In an MTMM experiment performed in the European Social Survey, the reliability coefficient of this particular question was also estimated directly from data. [3] These estimates from an actual MTMM experiment can be compared to the SQP predictions shown in Figure 5. In this MTMM experiment the reliability coefficient of this version of the question was estimated as 0.763 and the method effect at 0.038. Both are close to the predictions of these numbers obtained with SQP.

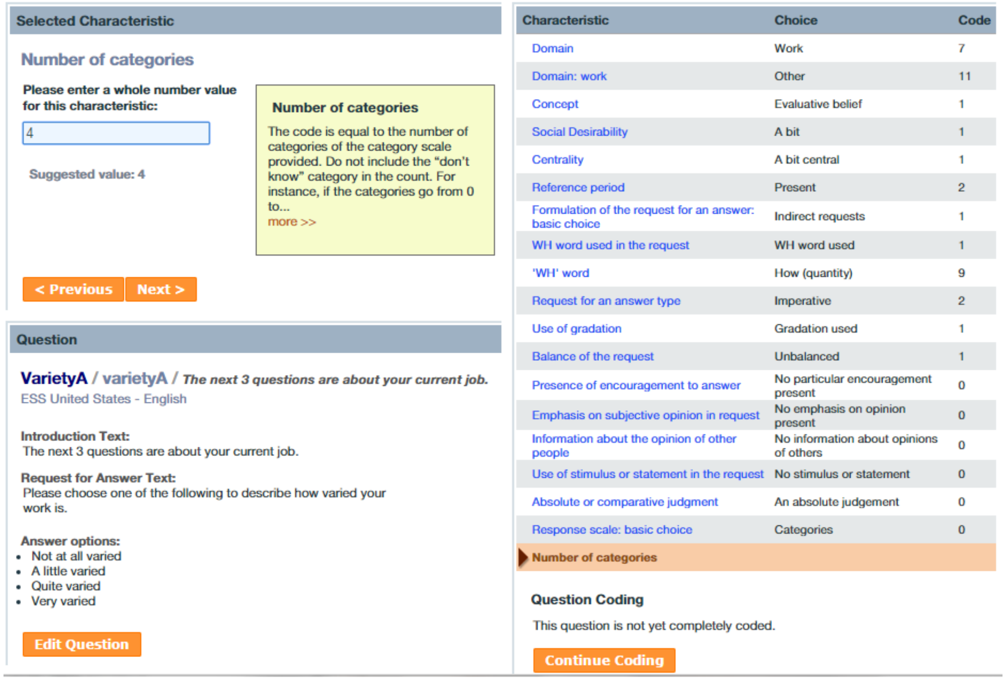
Finally, a tentative feature of SQP is that suggestions for potential improvement of the question are given. This is done by examining the “what-if” prediction that would be obtained from the random forest if one characteristic were coded differently. Figure 6 shows the suggestions made

[3] Program input and output for the MTMM analysis can be found at http://github.com/daob/ess-research/blob/master/input/mplus/Job/jobmtmm.out

by SQP 2.0: if the phrasing were simpler, in the sense of using fewer syllables per word and fewer words, the question would be predicted to have a higher quality. It is difficult to see how the question's phrasing (see Figure 1), which is already very simple, could be made even simpler. What could be changed is the ”scale correspondence”. This is the degree to which the numbers with which the answer options are labeled correspond to the meaning of the labels. In version A of the question, the labels are not numbered at all, so this correspondence has been coded as “low”. By introducing numbers 0, 1, 2, and 3 to go with the labels “not at all”, “a little”, “quite” and “very”, the scale correspondence could be coded as “high” and the predicted quality would improve somewhat.

This process could in principle be repeated until the question is thought to be of “acceptable” quality, or no further sensible improvements can be made. However, note that there may be good reasons *not* to make a possible suggested improvement when such an “improvement” does not make sense in the broader context of the questionnaire. Furthermore, note that since the meta-analysis does not directly address causality, there is no guarantee that this improvement in quality after changing the question will actually be realized. Addressing the causality of these changes remains a topic open for future research.

SQP should be placed in the much wider context of questionnaire science. For example, the meta-analysis finds that complicated phrasings are bad for reliability, something that others have also suggested and found (see Graesser et al., 2006). But additional explanations can also clarify meaning and narrow the range of possible interpretations a question has, reducing error (Fowler, 1992; Holbrook et al., 2006). This serves as a small demonstration that much more work needs to be done to synthesize the literature than could be achieved in this book chapter.



Selected Characteristic

Number of categories

Please enter a whole number value for this characteristic:

4

Suggested value: 4

Number of categories

The code is equal to the number of categories of the category scale provided. Do not include the "don't know" category in the count. For instance, if the categories go from 0 to...

Question

VarietyA / varietyA / *The next 3 questions are about your current job.*

ESS United States - English

Introduction Text:
The next 3 questions are about your current job.

Request for Answer Text:
Please choose one of the following to describe how varied your work is.

Answer options:

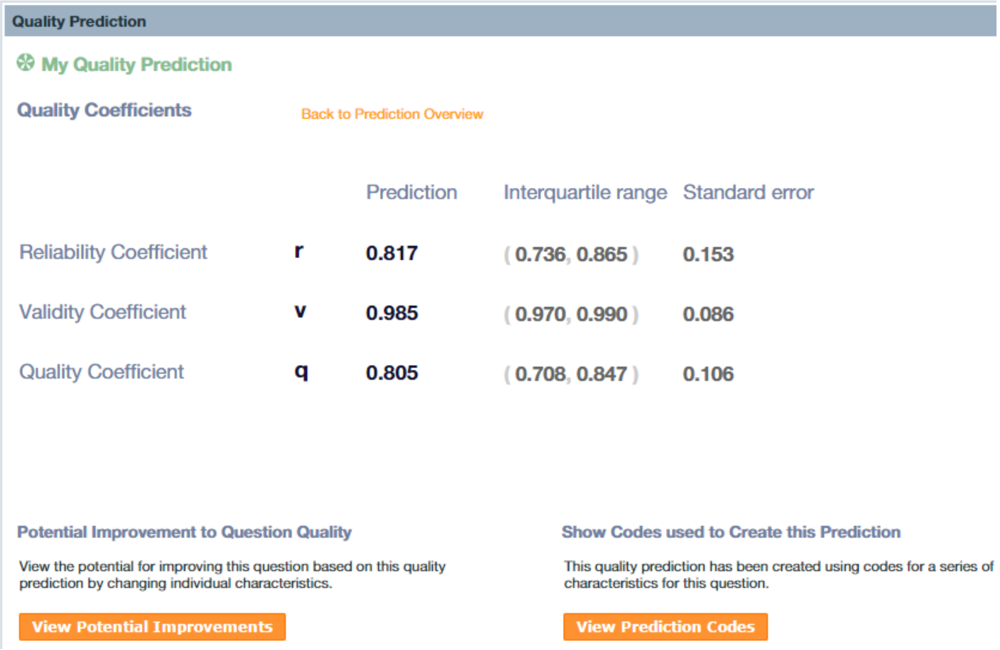
- Not at all varied
- A little varied
- Quite varied
- Very varied

Characteristic	Choice	Code
Domain	Work	7
Domain: work	Other	11
Concept	Evaluative belief	1
Social Desirability	A bit	1
Centrality	A bit central	1
Reference period	Present	2
Formulation of the request for an answer: basic choice	Indirect requests	1
WH word used in the request	WH word used	1
'WH' word	How (quantity)	9
Request for an answer type	Imperative	2
Use of gradation	Gradation used	1
Balance of the request	Unbalanced	1
Presence of encouragement to answer	No particular encouragement present	0
Emphasis on subjective opinion in request	No emphasis on opinion present	0
Information about the opinion of other people	No information about opinions of others	0
Use of stimulus or statement in the request	No stimulus or statement	0
Absolute or comparative judgment	An absolute judgement	0
Response scale: basic choice	Categories	0
▶ Number of categories		

Question Coding

This question is not yet completely coded.

Figure 4: Coding the characteristics of the questions in the system. More information on their precise meaning is given with each characteristic.



Quality Prediction

My Quality Prediction

Quality Coefficients

		Prediction	Interquartile range	Standard error
Reliability Coefficient	r	0.817	(0.736, 0.865)	0.153
Validity Coefficient	v	0.985	(0.970, 0.990)	0.086
Quality Coefficient	q	0.805	(0.708, 0.847)	0.106

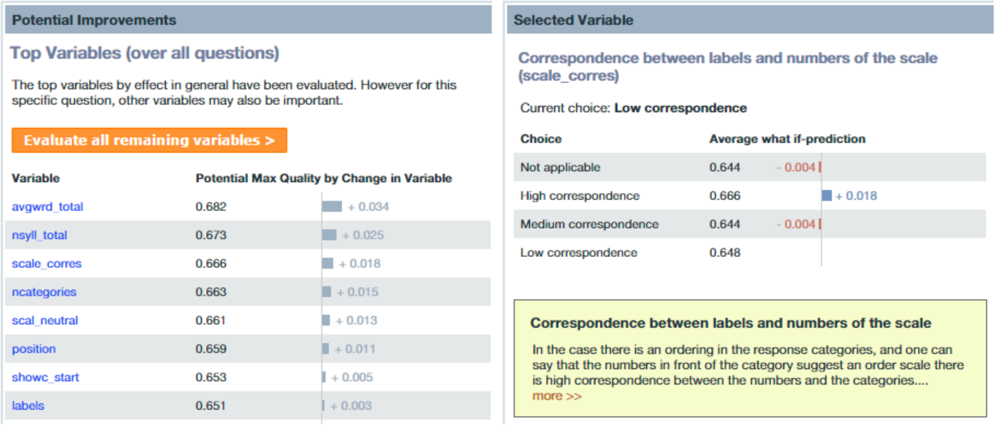
Potential Improvement to Question Quality

View the potential for improving this question based on this quality prediction by changing individual characteristics.

Show Codes used to Create this Prediction

This quality prediction has been created using codes for a series of characteristics for this question.

Figure 5: When the coding is complete, a prediction of the MTMM reliability and "validity" (complement of method effect) coefficients is given, together with the uncertainty about these predictions.



Potential Improvements

Top Variables (over all questions)

The top variables by effect in general have been evaluated. However for this specific question, other variables may also be important.

Variable	Potential Max Quality by Change in Variable	
avgwrd_total	0.682	+ 0.034
nsyll_total	0.673	+ 0.025
scale_corres	0.666	+ 0.018
ncategories	0.663	+ 0.015
scal_neutral	0.661	+ 0.013
position	0.659	+ 0.011
showc_start	0.653	+ 0.005
labels	0.651	+ 0.003

Selected Variable

Correspondence between labels and numbers of the scale (scale_corres)

Current choice: **Low correspondence**

Choice	Average what if-prediction	
Not applicable	0.644	- 0.004
High correspondence	0.666	+ 0.018
Medium correspondence	0.644	- 0.004
Low correspondence	0.648	

Correspondence between labels and numbers of the scale

In the case there is an ordering in the response categories, and one can say that the numbers in front of the category suggest an order scale there is high correspondence between the numbers and the categories....
more >>

Figure 6: SQP can look into its database of experiments to examine the differences in prediction that would occur if one aspect of the question were changed. The above suggests creating numbers to correspond with the labels might help.

5 Conclusion

The quest continues. We are far from understanding everything about how to ask the best possible questions, but can see that the only road to such knowledge is well-developed cognitive theory, careful empirical observation and experiment, and systematic synthesis of the body of knowledge. Steps on this road are taken in almost every issue of journals such as *Public Opinion Quarterly*, *Survey Research Methods*, and *Journal of Survey Statistics and Methodology*. Neither these individual steps, nor SQP, nor any textbook can give the definitive final word on questionnaire science. But all of these can help the researcher do better research, keeping in mind this chapter's counsels:

- We make a bewildering array of choices every time we formulate a survey question;
- Our personal experience does not guarantee knowledge about the optimal choices;
- Experts often have good advice to offer but are not exempt from the human tendency to overgeneralize;
- What is considered "best practice" differs over people and organizations, and may not correspond to actual best practice as observed in experiments.

In conclusion: always ask for the evidence. There may be plenty of it, or there may be little. Both cases offer an exciting chance to learn more about the science of surveys.

The future The year of this writing marks the 200th anniversary of the invention of a revolutionary new human measurement instrument. In 1816, René Théophile Hyacinthe Laennec, a young physician from a remote provincial town in France, found himself practicing in Paris. When a Parisian young lady entered his practice with heart problems, the modest young doctor hesitated to put his ear directly on her breast, as was the usual practice. Instead, he rolled a piece of paper into a cylinder with which he could hear his patient's heartbeat "much more neatly and distinctly" than he ever had before (Laennec, 1819, pp. 8–9). This new measurement method, the stethoscope, replaced the older ones.

Today, Laennec's stethoscope remains ubiquitous. Newer methods, such as x-rays and MRI, have not replaced it, but have complemented it. After all, a measurement method that is practical, fast, and cost-effective is hard to replace. The survey question is such a method in the social sphere. It therefore seems unlikely that newer measurement methods will fully replace the survey question in the foreseeable future. However, survey researchers and other students of human opinion and behavior should ponder the possible ways in which other measurements can be used to complement surveys. Furthermore, as argued in this chapter, the survey question still warrants improvement using modern methods of investigation. I will briefly elaborate on these two points below.

First, it is clear that the questionnaire is experiencing competition from other measurement instruments, old and new. Implicit association tests (Greenwald et al., 1998), for example, intend to measure prejudice with reaction times; fMRI and other brain imaging techniques show how the brain reacts to certain stimuli (Raichle and Mintun, 2006); genome-wide genetic sequencing has become feasible (Visscher et al., 2012); and data from companies' and governments' administrative registers provides some of the information we are after through record linkage (Wallgren and Wallgren, 2007). The use of everyday technology to measure human behavior is also becoming more popular. Monitoring smartphone usage with an app may be a better measure of smartphone usage than a questionnaire (Révilla et al., 2016); monitor the GPS in peoples' cars a better measure of their movements during the day (Cui et al., 2015); and Facebook (an online social network application from the early 21st century) "likes" strongly correlate with various personal characteristics (Kosinski et al., 2013).

All of these other measurement instruments are sometimes touted as being more "objective". I personally believe that this is not a helpful way to think about measurement (see also Couper, 2013). As we have seen, answers to questions have their biases and unreliabilities. But so do fMRI (Ramsey et al., 2010), GWAS studies (Visscher et al., 2012), administrative registers (Groen, 2012; Bakker and Daas, 2012; Kreuter and Peng, 2014), and "big data" such as Facebook posts or monitoring studies (Manovich, 2011; Fan et al., 2014). Furthermore, validity is often an issue with such measures: what if we were not interested in the person's movements and internet use, but in their political opinions, their desire to have children, or the people they fall in love with?

A more helpful way of thinking about these other instruments is as attempting to measure the same things that survey questions intend to measure. Which is the the best way of doing that, or whether perhaps several ways should be combined to obtain the best picture, is then an empirical matter that pertains to a particular research question. For example, Révilla et al. (2016) claimed that smartphone monitoring is better for measuring the amount of internet usage on a person's phone – no more, no less. Scientific experiments should then be used in the same way that we have been using them to look at the quality of survey measures alone. In short, no single measurement method is perfect. Instead, social researchers would do well to take a page out the medical practitioners' book and use a variety of measurement methods, old and new, cheap and expensive, and more or less reliable, valid, and comparable (Oberski, 2012), to zero in on the phenomenon being studied.

Aside from the inevitable opportunities and challenges afforded by the combination of surveys with other types of data, the survey question itself still warrants considerable improvement. This has been the topic of the current chapter, and SQP discussed as one attempt at such an improvement. However, this attempt is, of necessity, limited in scope and application. First, it has been applied only to a subset of questions, to specific groups of people, in a subset of countries, languages, and settings, during a particular time period. Second, it is only as good as the method used to measure the quality of survey questions, the MTMM experiment in this case. Third, it accounts for only certain aspects of the survey process and question characteristics. While the SQP project made every effort to widen its scope in each of these aspects, and does so over an impressive range of countries, settings, questions, and so forth, no project can cover every conceivable angle. Therefore, I see SQP's general philosophy, contributed by its fathers Frank Andrews and Willem Saris, as one of its most important contributions to the future of social research: that social measurement can be investigated scientifically.

In my ideal future, the Andrews-Saris approach to social research would become standard across the social sciences. Any way of measuring opinions, behavior, or characteristics of people would be studied by experiment. and the experiments summarized by meta-analyses that would be used to determine the best way to move forward. An example of a recent meta-analysis relating to nonresponse rather than measurement error is Medway and Fulton (2012). To ensure that such meta-analyses afford an appropriate picture of scientific evidence, we would also take into account lessons about the appropriate way to conduct science that are being learned in the emerging field of "meta-research"[4]. In particular, in addition to all the usual considerations for conducting good research, all conducted experiments should be published (Ioannidis, 2005), and preferably preregistered (Wagenmakers et al., 2012), conducted collaboratively ("copiloted"; Wicherts, 2011), and fully open and reproducible (Peng, 2011). When we all join in this effort, questionnaire science in particular, and the investigation of human opinion and behavior in general, will experience a huge leap forward.

References

Alwin, D. (2007). *Margins of error: a study of reliability in survey measurement*. Wiley-Interscience, New York.

Alwin, D. (2011). Evaluating the reliability and validity of survey interview data using the MTMM approach. In Madans, J., Miller, K., Maitland, A., and Willis, G., editors, *Question Evaluation Methods: Contributing to the Science of Data Quality*, pages 263–293. Wiley Online Library, New York.

Alwin, D. F. and Krosnick, J. A. (1991). The reliability of survey attitude measurement the influence of question and respondent attributes. *Sociological Methods & Research*, 20(1):139–181.

Andrews, F. (1984). Construct validity and error components of survey measures: A structural modeling approach. *Public opinion quarterly*, 48(2):409–442.

Bakker, B. F. and Daas, P. J. (2012). Methodological challenges of register-based research. *Statistica Neerlandica*, 66(1):2–7.

[4] See, for example http://metrics.stanford.edu/ and http://www.bitss.org/

Bradburn, N. M., Wansink, B., and Sudman, S. (2004). *Asking questions: the definitive guide to questionnaire design–for market research, political polls, and social and health questionnaires*. Jossey-Bass, San Francisco, rev. ed edition.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Campbell, D. and Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol Bull*, 56:81–105.

Carroll, R., Ruppert, D., Stefanski, L., and Crainiceanu, C. (2006). *Measurement error in nonlinear models: a modern perspective*, volume 105. Chapman & Hall/CRC Monographs on Statistics & Applied Probability, Boca Raton, FL.

Cernat, A. and Oberski, D. L. (2017). Extending the within-persons experimental design: The multitrait-multierror (MTME) approach. In Lavrakas, P. J., editor, *Experimental Methods in Survey Research*. John Wiley & Sons, New York.

Couper, M. P. (2013). Is the sky falling? new technology, changing media, and the future of surveys. In *Survey Research Methods*, volume 7, pages 145–156.

Cui, J., Liu, F., Hu, J., Janssens, D., Wets, G., and Cools, M. (2015). Identifying mismatch between urban travel demand and transport network services using gps data: A case study in the fast growing chinese city of harbin. *Neurocomputing*.

Dijkstra, W. and Smit, J. H. (1999). *Onderzoek met vragenlijsten: een praktische handleiding [Survey research: a practical guide]*. VU University Press, Amsterdam.

Dillman, D. A. (2011). *Mail and Internet surveys: The tailored design method–2007 Update with new Internet, visual, and mixed-mode guide*. John Wiley & Sons, New York.

Fan, J., Han, F., and Liu, H. (2014). Challenges of big data analysis. *National science review*, 1(2):293–314.

Fink, A. (2009). *How to conduct surveys: a step-by-step guide*. Sage, Los Angeles, 4th ed edition.

Folz, D. H. (1996). *Survey research for public administration*. Sage, Los Angeles.

Fowler, F. J. (1992). How unclear terms affect survey data. *Public Opinion Quarterly*, 56(2):218–231.

Fowler, F. J. (2014). *Survey research methods*. Sage, Los Angeles.

Fuller, W. (1987). *Measurement error models*. John Wiley & Sons, New York.

Graesser, A. C., Cai, Z., Louwerse, M. M., and Daniel, F. (2006). Question understanding aid (quaid) a web facility that tests question comprehensibility. *Public Opinion Quarterly*, 70(1):3–22.

Greenwald, A. G., McGhee, D. E., and Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464.

Groen, J. A. (2012). Sources of error in survey and administrative data: The importance of reporting procedures. *Journal of Official Statistics (JOS)*, 28(2).

Hagenaars, J. A. P. (1990). *Categorical longitudinal data: Log-linear panel, trend, and cohort analysis*. Sage, Newbury Park.

Heise, D. and Bohrnstedt, G. (1970). Validity, invalidity, and reliability. *Sociological Methodology*, 2:104–129.

Holbrook, A., Cho, Y. I., and Johnson, T. (2006). The impact of question and respondent characteristics on comprehension and mapping difficulties. *Public Opinion Quarterly*, 70(4):565–595.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Med*, 2(8):e124.

Költringer, R. (1995). Measurement quality in Austrian personal interview surveys. In Saris, W. and Münnich, A., editors, *The multitrait-multimethod approach to evaluate measurement instruments*, pages 207—225. Eötvös University Press, Budapest.

Kosinski, M., Stillwell, D., and Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805.

Kreuter, F. and Peng, R. D. (2014). Extracting information from big data: Issues of measurement, inference and linkage. In *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, page 257. Cambridge University Press.

Krosnick, J. (2009). Keynote lecture: Agree-disagree questions. In *Meeting of the European Survey Research Association*.

Krosnick, J. and Fabrigrar, L. (2001). *Designing questionnaires to measure attitudes*. Oxford University Press.

Laennec, R. T. H. (1819). *Traité de l'auscultation médiate, et des maladies des poumons et du coeur*, volume 1. J.-A. Brosson et J.-S. Chaudé libraires, Paris, Rue Pierre Sarrasin 9.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*.

Lord, F. M. and Novick, M. R. (1968). *Statistical theories of mental scores*. Addison–Wesley, Reading.

Madans, J., Miller, K., Maitland, A., and Willis, G. (2011). *Question Evaluation Methods: Contributing to the Science of Data Quality*. Wiley, New York.

Manovich, L. (2011). Trending: The promises and the challenges of big social data. *Debates in the digital humanities*, 2:460–475.

Medway, R. L. and Fulton, J. (2012). When more gets you less: a meta-analysis of the effect of concurrent web options on mail survey response rates. *Public opinion quarterly*, page nfs047.

Narayan, S. and Krosnick, J. A. (1996). Education moderates some response effects in attitude measurement. *Public Opinion Quarterly*, 60(1):58–88.

Netemeyer, R. G., Haws, K. L., and Bearden, W. O. (2011). *Handbook of marketing scales: multi-item measures for marketing and consumer behavior research*. Sage, Los Angeles, 3rd edition.

Oberski, D. (2012). Comparability of survey measurements. In Gideon, L., editor, *Handbook of Survey Methodology for the Social Sciences*, pages 477–498. Springer-Verlag, New York.

Oberski, D., Saris, W. E., and Kuipers, S. (2004). SQP: survey quality predictor.

Payne, S. L. (1951). *The art of asking questions*. Princeton U. Press, Oxford, UK.

Peng, R. D. (2011). Reproducible research in computational science. *Science*, 334(6060):1226.

Raichle, M. E. and Mintun, M. A. (2006). Brain work and brain imaging. *Annu. Rev. Neurosci.*, 29:449–476.

Ramsey, J. D., Hanson, S. J., Hanson, C., Halchenko, Y. O., Poldrack, R. A., and Glymour, C. (2010). Six problems for causal inference from fmri. *Neuroimage*, 49(2):1545–1558.

Révilla, M., Ochoa, C., and Loewe, G. (2016). Using passive data from a meter to complement survey data in order to study online behavior. *Social Science Computer Review*.

Révilla, M. A. (2012a). Impact of the mode of data collection on the quality of answers to survey questions depending on respondent characteristics. *Bulletin de Méthodologie Sociologique*, 116:44–60.

Révilla, M. A. (2012b). Measurement invariance and quality of composite scores in a face-to-face and a web survey. *Survey Research Methods*, 7(1):17–28.

Révilla, M. A. and Saris, W. E. (2012). A comparison of the quality of questions in a face-to-face and a web survey. *International Journal of Public Opinion Research*, page eds007.

Révilla, M. A., Saris, W. E., and Krosnick, J. A. (2013). Choosing the number of categories in agree–disagree scales. *Sociological Methods & Research*, page 0049124113509605.

Saris, W. and Gallhofer, I. N. (2007a). *Design, evaluation, and analysis of questionnaires for survey research*. Wiley-Interscience, New York.

Saris, W. E., A, R. M., Krosnick, J., and Shaeffer, E. M. (2010). Comparing questions with agree/disagree response options to questions with item-specific response options. *Survey Research Methods*, 4(1):61–79.

Saris, W. E. and Andrews, F. M. (1991). Evaluation of measurement instruments using a structural modeling approach. In Biemer, P., Groves, R., Lyberg, L., Mathiowetz, N., and Sudman, S., editors, *Measurement errors in surveys*, pages 575–599. John Wiley & Sons, New York.

Saris, W. E. and Gallhofer, I. (2007b). Estimation of the effects of measurement characteristics on the quality of survey questions. *Survey Research Methods*, 1.

Saris, W. E., Oberski, D. L., Révilla, M., Rojas, D. Z., Lilleoja, L., Gallhofer, I., and Gruner, T. (2012). Final report about the project JRA3 as part of ESS infrastructure (SQP 2002-2011). Technical report, RECSM, Universitat Pompeu Fabra, Spain, Barcelona.

Scherpenzeel, A. (1995). *A question of quality. Evaluating survey questions by multitrait-multimethod studies*. Royal PTT Nederland NV, Amsterdam.

Schuman, H. and Presser, S. (1981). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Sage, Thousand Oaks, CA.

Selznick, G. J. and Steinberg, S. (1969). *The tenacity of prejudice: Anti-Semitism in contemporary America*. Harper & Row, Oxford, England.

Tourangeau, R., Rips, L., and Rasinski, K. (2000). *The psychology of survey response*. Cambridge Univ Press, Cambridge, United Kingdom.

Trabasso, T., Rollins, H., and Shaughnessy, E. (1971). Storage and verification stages in processing concepts. *Cognitive psychology*, 2(3):239–289.

Visscher, P. M., Brown, M. A., McCarthy, M. I., and Yang, J. (2012). Five years of gwas discovery. *The American Journal of Human Genetics*, 90(1):7–24.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L., and Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, 7(6):632–638.

Wallgren, A. and Wallgren, B. (2007). *Register-based statistics: administrative data for statistical purposes*. Wiley, New York.

Wicherts, J. M. (2011). Psychology must learn a lesson from fraud case. *Nature*, 480:7.

Wiley, D. and Wiley, J. A. (1970). The estimation of measurement error in panel data. *American Sociological Review*, 35(1):112–117.

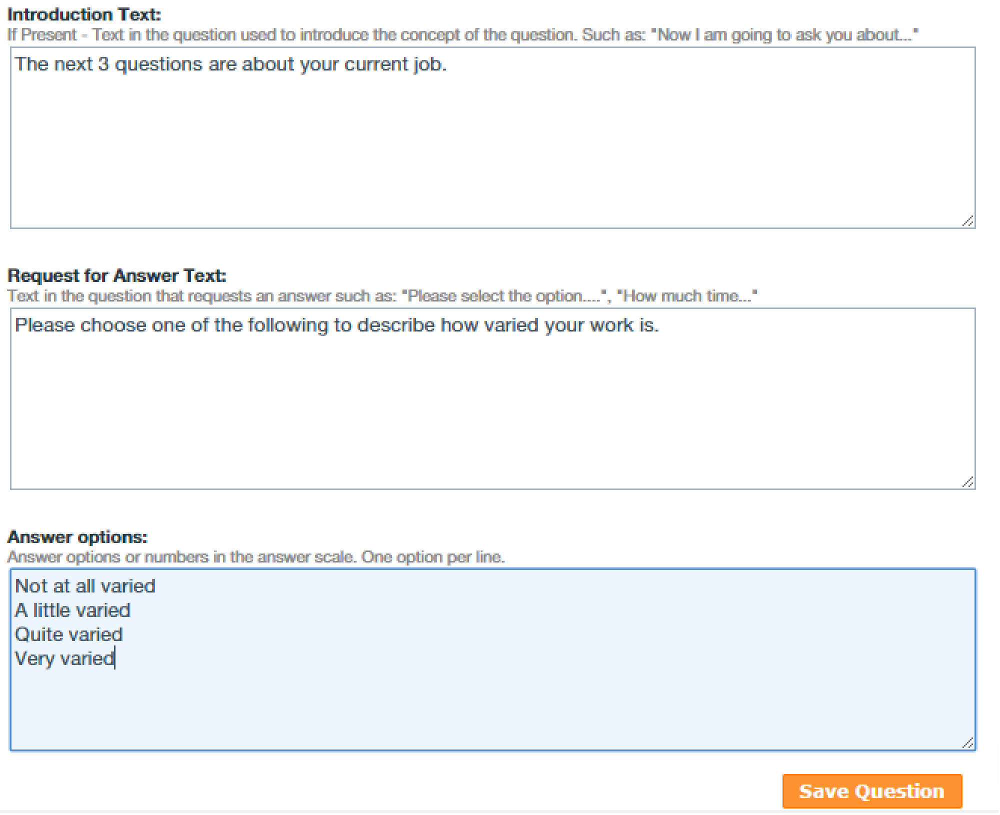
A Full list of choices made in SQP 2.0

Below is the full list of choices I made for the characteristics of the ”job variety” question in Figure 1 using SQP 2.0 (`http://sqp.upf.edu/`). Further explanations about the precise meaning of these codes can be found while coding on the website as well as in Saris and Gallhofer (2007a).

Characteristic	Choice	Code
Domain	Work	7
Domain: work	Other	11
Concept	Evaluative belief	1
Social Desirability	A bit	1
Centrality	A bit central	1
Reference period	Present	2
Formulation of the request for an answer: basic choice	Indirect requests	1
WH word used in the request	WH word used	1
‘WH’ word	How (quantity)	9
Request for an answer type	Imperative	2
Use of gradation	Gradation used	1
Balance of the request	Unbalanced	1
Presence of encouragement to answer	No particular encouragement present	0
Emphasis on subjective opinion in request	No emphasis on opinion present	0
Information about the opinion of other people	No information about opinions of others	0
Use of stimulus or statement in the request	No stimulus or statement	0
Absolute or comparative judgment	An absolute judgement	0
Response scale: basic choice	Categories	0
Number of categories	4	4
Labels of categories	Fully labelled	3
Labels with long or short text	Short text	0
Order of the labels	First label negative or not applicable	1
Correspondence between labels and numbers of the scale	Low correspondence	3
Theoretical range of the scale bipolar/unipolar	Theoretically unipolar	0
Number of fixed reference points	0	0
Don’t know option	DK option not present	3

Interviewer instruction	Absent	0
Respondent instruction	Present	1
Extra motivation, info or definition available?	Absent	0
Introduction available?	Available	1
Number of sentences in introduction	1	1
Number of words in introduction	9	9
Number of subordinated clauses in introduction	0	0
Request present in the introduction	Request not present	0
Number of sentences in the request	1	1
Number of words in request	13	13
Total number of nouns in request for an answer	2	2
Total number of abstract nouns in request for an answer	1	1
Total number of syllables in request	17	17
Number of subordinate clauses in request	0	0
Number of syllables in answer scale	16	16
Total number of nouns in answer scale	0	0
Total number of abstract nouns in answer scale	0	0
Show card used	Showcard not used	0
Computer assisted	Yes	1
Interviewer	Yes	1
Visual presentation	Oral	0
Position	50	50

B SQP screenshots



Introduction Text:
If Present - Text in the question used to introduce the concept of the question. Such as: "Now I am going to ask you about..."

The next 3 questions are about your current job.

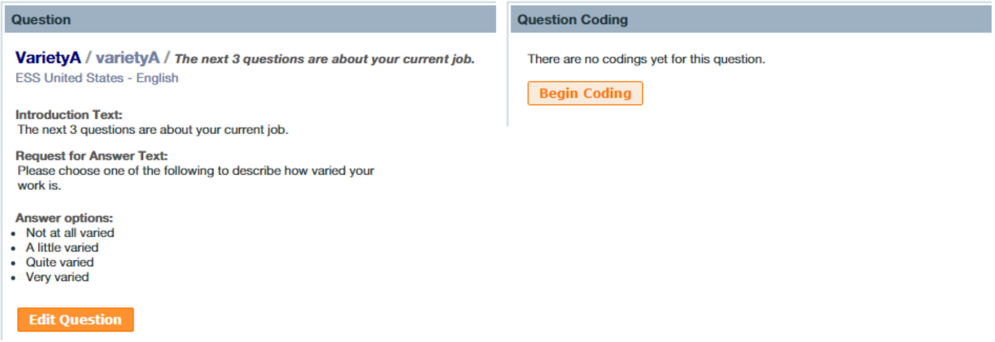
Request for Answer Text:
Text in the question that requests an answer such as: "Please select the option....", "How much time..."

Please choose one of the following to describe how varied your work is.

Answer options:
Answer options or numbers in the answer scale. One option per line.

Not at all varied
A little varied
Quite varied
Very varied

Figure 7: Entering the ”job variety” question into the SQP system.



Question

VarietyA / varietyA / ***The next 3 questions are about your current job.***
ESS United States - English

Introduction Text:
The next 3 questions are about your current job.

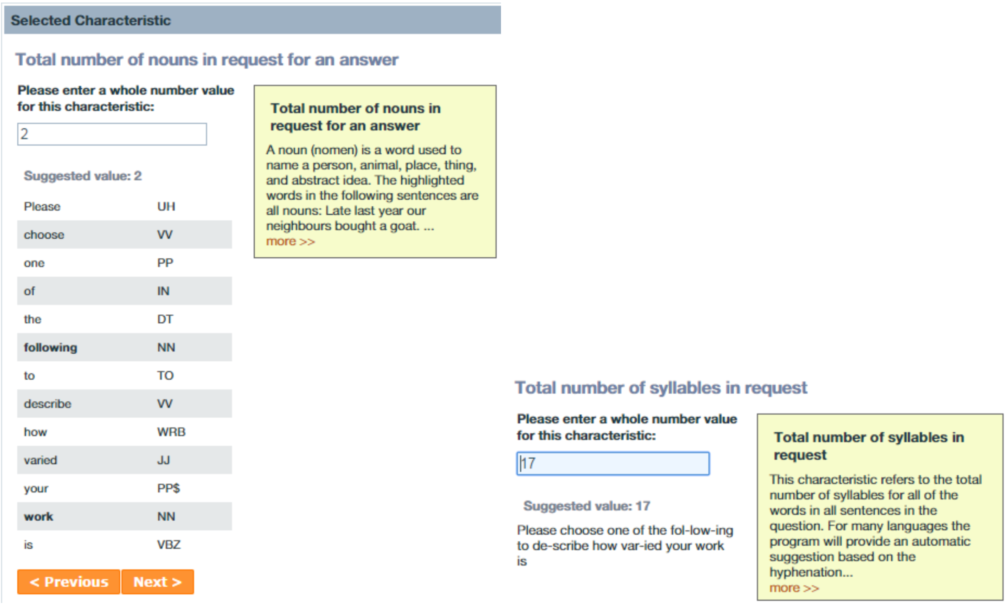
Request for Answer Text:
Please choose one of the following to describe how varied your work is.

Answer options:
- Not at all varied
- A little varied
- Quite varied
- Very varied

Question Coding

There are no codings yet for this question.

Figure 8: The SQP opening screen to begin coding the question.



Selected Characteristic

Total number of nouns in request for an answer

Please enter a whole number value for this characteristic:

2

Suggested value: 2

Please	UH
choose	VV
one	PP
of	IN
the	DT
following	NN
to	TO
describe	VV
how	WRB
varied	JJ
your	PP$
work	NN
is	VBZ

Total number of nouns in request for an answer

A noun (nomen) is a word used to name a person, animal, place, thing, and abstract idea. The highlighted words in the following sentences are all nouns: Late last year our neighbours bought a goat. ...
more >>

Total number of syllables in request

Please enter a whole number value for this characteristic:

17

Suggested value: 17

Please choose one of the fol-low-ing to de-scribe how var-ied your work is

Total number of syllables in request

This characteristic refers to the total number of syllables for all of the words in all sentences in the question. For many languages the program will provide an automatic suggestion based on the hyphenation...
more >>

Figure 9: Some characteristics, such as the number of nouns and syllables, are detected automatically using natural language processing techniques. Others must be coded by hand.