

WHEN OFTEN BECOMES ALWAYS, AND SOMETIMES BECOMES NEVER: MISCOMPREHENSION IN SURVEYS

BEN HARDY

The Open University Business School
Walton Hall
Milton Keynes MK7 6AA
United Kingdom

LUCY R FORD

Saint Joseph's University

INTRODUCTION

Much of social science involves soliciting the opinions of others, often through the use of surveys. There are a number of different survey formats but a commonly used approach is to provide a stimulus question which is coupled to a limited number of response options. This approach has several benefits. First, the structured format presents the stimulus to each respondent in the same format, reducing inconsistency of stimulus presentation. Second, the standardized format of responses facilitates processing of responses. Third, surveys in this format act as a transduction point where the verbal stimulus of the question is converted to a numerical response through the allocation of ordinal values.

This whole enterprise, however, relies on an unbroken chain of comprehension from the mind of the researcher, through the survey, to the mind of the respondent, and back again. Miscomprehension at any stage in this process introduces error which, in turn, impacts the quality and utility of the measure. It is not enough, however, that respondents understand the items. They must understand the survey items in the same way as the researcher; otherwise they are not responding to the construct that the researcher intended.

The advantages of surveys outlined above, which convert verbal (analog) stimuli to numerical (digital) output are also their weakness, as a simple numerical output has the potential to obscure any miscomprehension of the instruction, question or response. Individuals may 'strongly agree' with a statement but not necessarily 'strongly agree' with the same thing as other respondents. This difference is unlikely to be apparent statistically. But if instruments do not accurately measure the phenomena of interest then the research founded upon those instruments is of diminished value.

This paper examines miscomprehension and variance in interpretation of survey item through two empirical studies. We find considerable evidence that, whilst respondents understand survey items, they do not all do so in the same way. This variance in comprehension cannot readily be detected statistically, and yet is a significant threat to construct validity, casting doubt on whether the instruments are tapping into the same construct for all respondents.

BACKGROUND TO THE RESEARCH

A brief review of survey development

The process of scale and survey development begins with item development (Hinkin, 1998), a stage which is often not accorded sufficient emphasis as researchers “often ‘throw together’ or ‘dredge up’ items and assume they constitute a suitable scale” (DeVellis, 2003: 11). Hinkin (1998) advocates a rigorous process of item development to ensure that parsimonious questions are written in language familiar to the putative respondents and attempts are made to examine construct validity using multiple samples and techniques such as exploratory and confirmatory factor analysis.

Researchers may instead choose to use published measures, believing that because they have successfully passed through the peer review process, they are beyond reproach. Ford and Scandura (2007) examined a compilation of organizational measures (Fields, 2002) and found that the majority of scales contained one or more threats to construct validity, suggesting that published measures are not without flaws.

Using a published measure also involves disembedding it from its original context, potentially increasing the risk of error. Comprehension is a function of both syntax and context, and so the same syntax may be comprehended differently in a different context. The importance of linguistics in item interpretation has not been widely discussed although there are some existing sources which do address the issue (e.g. Schwarz, 1999; Tourangeau, 2000).

A brief review of relevant linguistic theory

Surveys hinge on comprehension, yet linguistic theory suggests that the indeterminacy of language means that this comprehension is inevitably imperfect. Researchers in the pragmatic tradition of semantics make a clear distinction between what is said, and the context (both social and linguistic) in which it is said; the one influencing the other. The interplay between the contextual and literal was articulated by Grice (1975; 1978) and subsequently modified and extended by other authors (e.g. Horn, 1984; Sperber, 1986; Levinson, 2000; Jaszczolt, 2005). The principle underpinning this field is that the bare syntactic skeleton of a sentence is fleshed out by the respondent with contextual, environmental and experiential information to produce an understanding of what the question is asking. This process is neither uniform nor predictable, varying across individuals and situations.

Types of error

Threats to comprehension, and hence validity, fall into two basic categories: instructional and interpretive. Interpretive errors can then be further broken down into two categories commonly used within the psycholinguistic literature (see Hernandez, 2001; Faust, 2002; Davis, 2005). One concerns the comprehension of the full sentence, or sentential comprehension, and the other the comprehension of individual words, or lexical comprehension.

Instructional miscomprehension

This is the most readily understood source of error. In this case the respondent does not read / follow the instructions for completing the survey, or they misunderstand the instructions (Tourangeau, 2000).

Sentential miscomprehension

This form of error concerns how respondents understand an item in its entirety. When respondents attempt to comprehend an item they do not undertake a simple banaisic process whereby they respond to the literal meaning of the sentence. Rather they make decisions, sometimes consciously, sometimes unconsciously, about what the question means.

Sentences can be enriched, where respondents go beyond the syntax of the sentence, or they can be depleted, where respondents impoverish the meaning of the sentence.

Lexical miscomprehension

This form of miscomprehension concerns the meaning of the words themselves. One person's definition of the word does not necessarily accurately map onto that of another's, for a variety of educational, cultural, social, contextual or gender-specific reasons. The word 'satisfaction', for example, has two historical meanings (Oxford English Dictionary, 2008: 502). One is 'with reference to desires or feelings' (pleasurable) and the other 'with reference to obligations' (transactional).

These three forms of miscomprehension, instructional, sentential and lexical, have the potential to introduce considerable error into the measurement process.

STUDY 1– RESPONDENT INTERPRETATIONS OF PUBLISHED SCALES

Method and sample

In order to explore the three forms of miscomprehension outlined above, a survey was conducted. Respondents were asked to write down what they thought the survey question was asking, imagining that they were explaining the item to someone who does not speak English and attempting to convey the true meaning of the item. In addition to describing the meaning, participants also responded to the items conventionally, using a Likert scale.

The survey questions we used were drawn from Dail Fields' book (2002). We used two job satisfaction scales (Tsui, 1992; Agho, 1992) and one procedural justice scale (Sweeney, 1993). The scales selected were published in respected journals and have been widely used, and Ford and Scandura (2007) found them to be free from threats to construct validity.

Respondents were from three groups: native speakers of British English (n=41), native speakers of American English (n=40) and members of the RMNet listserv (n=34).

Results

Cronbach's alpha scores were acceptable, ranging from 0.79 to 0.91. In exploratory factor analyses the scales loaded appropriately.

Instructional miscomprehension

Respondents were instructed to write down what they thought the question was asking, and not to attempt to answer the question. Nonetheless eight respondents consistently answered the questions instead of describing them (British 3/41; American 2/40; RMNet 3/34). The

RMNet group also provided additional responses which suggested that they had not read the question. Overall 17/155 (15%) respondents provided answers which suggested that they had not read the instructions properly.

Sentential miscomprehension

Two processes of sentential miscomprehension were examined: enrichment and depletion. Examination of the sample revealed evidence of both, with high levels of interrater reliability ($\kappa = 0.85-0.90$).

Depletion was particularly evident with the items from the procedural justice scale (Sweeney, 1993), where failure to apprehend the procedural element meant that questions such as 'How fair or unfair are the procedures used to determine pay rates?' were interpreted as 'How fair is your pay?' Every question in this scale showed this depletion with 21-31% of respondents neglecting the procedural element of the question. On average, across all items, 19.3/115 (17%) of respondents depleted the stimulus items.

Enrichment was far more common, which is understandable as linguistic theory suggests that individuals flesh out the basic syntax of the item in an attempt to understand it. The level of enrichment ranged from 21 to 55%. For example, the item 'I am satisfied with my job for the time being' (Agho et al., 1992) seemed to trigger an association with turnover intention. This enrichment was not uniform, however. For example one respondent interpreted this as 'I am currently happy with my job, but I may look for a new job in the future.' whilst another interpreted it as 'I'll be out of here at the first opportunity'. Both these respondents 'agreed' with this item. This demonstrates that numerical output from multiple item scales can mask considerable linguistic variance as two completely opposite interpretations had the same score (4/5).

The statistical impact of enrichment and depletion was assessed by comparing the enriched and depleted groups with the rest of the sample. No significant difference was detectable between either the depleted responses or the enriched ones and the rest of the sample.

Lexical miscomprehension

Lexical miscomprehension is the most difficult to detect, as the mental schema the individual draws on to define a word is unknown to the researcher. The two different senses of the word 'satisfaction' outlined above were used to examine lexical miscomprehension in eight of the items. An average of 38% of respondents used a pleasurable definition of satisfaction, 18% transactional and 16% neutral.

Another example of lexical miscomprehension was seen with the word "often" which was interpreted to mean anything from 33% of the time to 99%, a variation of 300%. The word 'most' resulted in similar variations in interpretation. In another vein, the comparators for the item 'I like my job better than the average worker does' were examined. Three categories of who represented the average worker emerged: the respondent's peers (21%), the general population (34%), and individuals in a similar job (4%, $\kappa = 0.88$). No statistical difference in response to the items was noted between any categories.

These results demonstrate evidence of all three types of miscomprehension, despite their being statistically undetectable. This suggests that there is considerable linguistic ambiguity within these scales which is potentially a significant but hidden source of error.

STUDY 2 – SELF-CLASSIFICATION INTO TYPES OF MISCOMPREHENSION

Methods and sample

The same scales as in Study 1 were presented in two different ways. The first was the conventional format with a Likert scale response. The second section presented respondents with the items from each of the three scales and a number of different response options, drawn from the open ended responses in Study 1. When assessing sentential miscomprehension, five different options were offered: A neutral interpretation of the original item i.e. a paraphrasing, an enriched item, a depleted item, an item which had been both enriched and depleted, and a statement that ‘None of the interpretations offered match my interpretation of the question/statement’.

To measure lexical miscomprehension we offered responses with differing interpretations of the word ‘satisfaction’ (as intended, as pleasurable, and as transactional). Lexical miscomprehension was also assessed in the three items with ambiguous modifying terms including often, most, and average worker.

The survey was administered through Qualtrics using a snowball sampling approach. 148 valid responses were received with most respondents being in the UK (61%) or US (26%).

Results

The numerical responses to the first portion of the survey were statistically analyzed. The scales were reliable ($\alpha=0.87-0.91$) and loaded appropriately in exploratory factor analysis.

Sentential miscomprehension

Over 90% of the respondents found that one of the four interpretations of the stimulus item offered (i.e. neutral, enriched, depleted or enriched and depleted) coincided with their interpretation of the stimulus item. If respondents had followed the strict syntax of the stimulus item then they should have selected the neutral response. Although this was the modal category and was, on average, selected 54% of the time (range 28-71%), enriched items were selected 21% of the time (range 3-36%), depleted 12% (range 2-28%), and both depleted and enriched 9% (range 4-16%). The data provide strong support for the findings of Study 1, although enrichment and depletion rates are somewhat lower. Nonetheless, these data suggest that sentential miscomprehension is widespread and a cause for concern, as it may threaten construct validity and the scores of respondents.

Lexical miscomprehension

Over 90% of respondents found that one of the options that they were presented with matched their own interpretation of the item. Analysis of respondents’ understanding of the word ‘satisfied’ suggests that, again, there was significant deviation from the neutral response. 59% of respondents selected the ‘neutral’ option (range 36-71%). 25% of respondents selected a pleasurable interpretation of the word ‘satisfied’ (range 15-33%) and 12% chose the transactional response (range 4-27%).

The word 'often' in the item 'I am often bored with my work' was interpreted as ranging from 33% of the time to more than 75% of the time. The word "most" and the term "average worker" were equally interpreted broadly.

These data present strong evidence that individuals interpret words differently. Some regard being 'satisfied' as a pleasurable experience whereas others regard it as a transactional one. Modifying words such as 'often' are similarly open to interpretation as individuals draw on different mental schema to define them.

DISCUSSION

The results of these two studies provide strong evidence for three types of miscomprehension: instructional, sentential and lexical. With regard to instructional miscomprehension, it is clear that respondents may not respond to instructions as the researcher intended. This suggests that some sort of detection system for failure to read instructions should be incorporated into all surveys. This would help identify respondents who have not read the instructions and/or questions and allow their responses to be segregated.

In terms of sentential miscomprehension, it appears that about 1/3 of the respondents are taking a very different survey to the one which the researcher thinks they are. This is likely to have both a direct and an indirect impact on validity, either by tapping into another construct as a result of either enrichment or depletion, or by creating collinearity with another construct within the model. Sentential miscomprehension, therefore, has the potential to introduce considerable error into survey results and this error is unlikely to be detectable using conventional approaches.

What is the impact of lexical miscomprehension? It is possible that some respondents may be tapping into different meanings when they interpret items than do others. This will have both primary effects on the scale itself and secondary effects on any model incorporating concepts similar to those being unconsciously tapped into by the respondent.

The findings of this study make uncomfortable reading for those of us involved in survey research. The reliability and validity of measurement instruments and surveys is of pivotal importance. If the measures either do not measure what is purported or do not do so accurately, then recommendations based upon research that employs these measures are likely to be flawed.

So what is to be done? Uniformity of instruction may reduce instructional miscomprehension, as may drawing attention to the instructions. Sentential miscomprehension may be reduced by grouping similar items together (to aid contextual interpretation) or underlining key words (e.g. 'process') to ensure that they are attended to. Lexical miscomprehension could be improved by eschewing vague words such as 'many', 'often' or 'sometimes'. Miscomprehension could be identified using Q-sort methodology (Schriesheim et al, 1993) to help measure the differences between individual judges, or Anderson and Gerbing's (1991) methods for substantive validity analysis, or cognitive interviewing (e.g. Beatty, 2007).

Survey research is a critical weapon in the social scientist's methodological armory. This paper is not intended to denigrate surveys as an information source or research tool but rather it seeks to draw the reader's attention to some of the linguistic problems which underlie surveys and to demonstrate the magnitude of effect of these problems. Attention to the potential methodological issues outlined in this paper should help produce better, more valid results which will, in turn, provide the basis for an improved understanding of social phenomena.

REFERENCES AVAILABLE FROM THE AUTHORS