Set Them Free: Improving Data Quality by Broadening the Interviewer’s Tasks

Giampietro Gobo

Received 18 August 2003; Accepted 9 May 2005

The paper deals with a highly controversial issue in survey data collection: the standardization of the interviewer’s behaviour during the interviewee’s selection of response alternatives. In the light of a large set of data drawn from several methodological studies published in the last 50 years, the author documents a counter-intuitive issue: (1) interviewer’s errors are of secondary importance and far smaller than respondents’ errors; and (2) in order to minimize respondent’s errors, we need to broaden the interviewer’s tasks. Focusing on the unsolved problem of multiple word meanings of response alternatives as a relevant part of response bias, the author argues that data quality can be achieved by entrusting to the interviewer a more active role. Of course, the aim of reducing respondent’s errors by broadening interviewer’s tasks will surely produce an increase in the interviewer’s effects on answers. However, the dilemma to be faced is which kind of errors we prefer (and are more useful) to minimize.

Introduction

A great deal of research has been conducted during the last 50 years, exploring cognitive aspects of the process of answering in surveys. Focusing on the interplay between cognitive processes and questionnaire design, several sources of bias have been highlighted: question wording, question form (acquiescence, balance, etc.), reaction to object, items curvilinearity, question order, and item order.

Drawing on this perspective, this article is devoted to one of the most important parts of the whole questionnaire: the response alternative. Schwarz and Hippler (1987)
have already shown that response alternatives are far from being a passive instrument collecting respondent’s behaviour, attitudes, or opinions only, because response alternatives play an ‘informative function’, affecting both respondents’ attitudes and recall processes. From a phenomenological point of view, the reflection on the response alternative is particularly significant because answers are, in some way, more important than questions: we perform a data analysis on answers not on questions. In other words, scaling is the technique which transforms a respondent’s attitude into a response, a (complex) opinion in a (simple) answer which receives a code and becomes a datum by means of the numerical matrix. The gap (big or small?) between attitude and response is still a problematic issue. Although this last problem might give rise to criticism of scaling as a rigid simplification of respondent’s experience and feelings, it still remains a vital, irreplaceable, and non-renounceable tool in reducing the complexity of information embedded in an attitude, opinion, or behaviour.

Concerning the response alternative, major findings discuss effects related to response order, open-ended vs. closed-ended response alternative, response alternative structure (rating vs. ranking; number of scale points or range of rating; verbal vs. numerical labels), and semantic vagueness of quantifiers. However, the semantics and pragmatics of response alternatives in standardized interviews have not yet been fully investigated. In addition, the standardization of interviewer’s behaviour, the traditional solution proposed to reduce response effects, does not succeed adequately the aim (which should be shared by every social scientist) of saving the (several?) meanings attributed by respondents to response alternatives.

To better understand the importance of this issue, it might be useful to return briefly to the construction of scaling and related response alternatives. First, Thurstone (1928) proposed a technique to measure attitudes. He suggested considering attitude as a *continuum* divisible in Equal Appearing Intervals, and his idea has been largely implemented in survey research practice. Later on, Likert (1932) presented another technique of scaling, still based on the same idea of equidistance between each scale points. Following this, scaling has been enriched by ‘Successive Intervals’ (Saffir, 1937) and ‘Own Categories’ (Sherif & Hovland, 1953), both derived from Thurstone’s proposal. These techniques measure quasi-cardinal variables, that is variables obtained by properties (e.g. beliefs, attitudes, opinions, and feelings), which survey researchers consider continuous but which cannot be ‘measured’ (Cicourel, 1964), since they lack a proper unit of measure (as is the case in physics for ‘time’ or ‘space’). This manipulation can be accepted under certain conditions:

1. the scale is ordinal, that is with a sequence (series) of ordered categories which must be perceived in the same way by all respondents;
2. the intervals among response alternatives (or scale points) are perceived as equal by all respondents;
3. the scale is uni-dimensional, that is each response alternative has one and only one meaning (semantics) for all respondents, and it is used (pragmatics) in the same way by them;
4. there is a natural correspondence between psychological and numerical intervals.
The first and second conditions involve mainly a cognitive problem. The third and fourth conditions concern a semantic/pragmatic phenomenon. I now develop this argument.

**Intervals Among Response Alternatives**

The existence of equal intervals between scale points has been questioned and criticized by Galtung (1967), Jordan (1965), Marradi (1980/1981), and Pawson (1982) because often respondents do not perceive as equal the intervals among response alternatives or scale points. This cognitive phenomenon has also been demonstrated by some research conducted in Italy: in Alessandria district by Amisano and Rinaldi (1988), in Naples by Buongiorno (1987), in Sicily by Cacciola and Marradi (1988), and in Florence district by Pampanin (1988).

Amisano and Rinaldi adopted the ‘scale of ideology’ invented by Blackburn and Mann (1975). This scale consists of the following seven response alternatives (or scale points): very strongly disagree/strongly disagree/mildly disagree/half and half/mildly agree/strongly agree/very strongly agree.

They used two different questionnaire administrations. In the first (1980), the response alternatives were located on a ruler. In the second (1984), the response alternatives were administrated in traditional form (that is, without the ruler), the form generally used in surveys.

In contrast to Stevens’ (1946) assumption that intervals among response alternatives are equal, using the Multiple Correspondence Analysis (MCA) Amisano and Rinaldi (1988) found that in the last administration (that is without the ruler) many respondents considered the three response alternatives concerning the area of disagreement: ‘very strongly disagree/strongly disagree/mildly disagree’, to be very close and even interchangeable. Similarly even the intervals between ‘mildly agree/strongly agree/very strongly agree’ were not equal: ‘mildly agree’ and ‘strongly agree’ were perceived as closer, while ‘very strongly agree’ was perceived as very much distant from the latter (figure 1).

The bias discussed above was noticed across the whole sample. This means that we are facing not a ‘social’ phenomenon, i.e. concerning a particular social category of people (based on gender, age, language, place of birth, education, religion, social class, profession, low-education etc.), but a ‘cognitive’ phenomenon (i.e. due to immediate perception) which applies across the sample.\(^2\)

Even more worrying were Buongiorno’s (1987) results drawn from a survey in Naples: many low-educated juvenile respondents considered ‘mildly agree’ a stronger evaluation than ‘very strongly agree’; that is the respondents systematically interchanged the meanings that researchers attributed to the response alternatives. In this way, the scale is not ordinal anymore.\(^3\) Cacciola and Marradi (1988) state that this phenomenon is widespread in the whole of southern Italy.

Applying the MCA to the data collected in 1980 (the first administration), Amisano and Rinaldi (1988) clearly show that the simultaneous use of response alternatives embedded in the ruler reduces response errors and represents a useful aid to improve respondents’ perception of intervals. In fact, the spatial analogy implied by the ruler
helps respondents to perceive as equal the intervals among response alternatives, as shown in figure 2.

In order to estimate the mean distance perceived between the response alternatives by respondents, the authors constructed a contingency table with the seven response alternatives on the columns and the nine items on the rows. Applying the MCA technique to this contingency table, they obtained the factors (see Table 1).

Amisano and Rinaldi (1988, p. 58) conclude that:

The spatial analogy has helped to stipulate a more effective cognitive agreement between researcher and respondents about the rules which govern the distance among response alternatives selected by respondents. Generally speaking the efficacy of spatial analogy could be attributed to the larger universality (intersubjectivity) of social conventions about physical distances rather than to conventions about semantic distances.

However, looking at this last graph, we can state that although the ruler has improved the perception and the communication between researcher and respondents, it has only reduced this bias without totally solving the problem because the intervals between the response alternatives in the disagreement area are still not equal, and individual idiosyncrasies (as reactions to item-wording, tendency to use only a part of the scale or the extreme response style) are not under control (Amisano & Rinaldi,
Figure 2. First two factors drawn in the 1980, with the survey with response alternatives located on a ruler. Legend — lambda: first axis 0.07 (72%); second axis 0.02 (17%). Source: Amisano and Rinaldi (1988, p. 56)

<table>
<thead>
<tr>
<th>Traditional values</th>
<th>Response alternatives</th>
<th>1980 (with ruler)</th>
<th>1984 (without ruler)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Very strongly disagree</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Strongly disagree</td>
<td>1.15</td>
<td>1.45</td>
</tr>
<tr>
<td>3</td>
<td>Mildly disagree</td>
<td>2.17</td>
<td>1.74</td>
</tr>
<tr>
<td>4</td>
<td>Half and half</td>
<td>3.65</td>
<td>3.29</td>
</tr>
<tr>
<td>5</td>
<td>Mildly agree</td>
<td>4.77</td>
<td>4.52</td>
</tr>
<tr>
<td>6</td>
<td>Strongly agree</td>
<td>5.74</td>
<td>5.16</td>
</tr>
<tr>
<td>7</td>
<td>Very strongly agree</td>
<td>7</td>
<td>7</td>
</tr>
</tbody>
</table>

Source: Amisano and Rinaldi (1988, p. 60).
In addition, the MCA technique does not uncover the meaning attributed by respondents to the response alternatives. So, we need to take a further step.

**Semantics of Response Alternatives**

Groves *et al.* (1992, p. 49) underline: ‘Whereas there has been much attention paid to the impact of question wording in the survey methodology literature, there have been only few attempts to examine perceived meaning ….’ The same criticism can be applied also to the response-wording literature. However, some researchers have found the existence of multiple word meanings of response alternatives phenomenon. For example, Hakel’s (1968) study replicated and achieved the same results as research conducted by Simpson (1944) on the perception of 20 terms (as ‘always’, ‘very often’, ‘often’, ‘never’, and so on) indicating degrees of frequency. Several researchers have pointed out the inexorable presence of multiple word meanings of response alternatives because of the communicative functions of quantifiers (Moxey & Sanford, 1992).

This phenomenon undermines the unidimensionality of the scale, the basis of comparability among respondents’ responses and, in a broad sense, the validity of research results. In order to investigate in detail this second bias, I conducted a socio-linguistic study concerning the multiple word meanings attributed by Italian respondents to four response alternatives commonly used in questionnaires to evaluate personal and work relations: *strongly satisfying*/*fairly satisfying*/*fairly unsatisfying*/*strongly unsatisfying*. Analysing 80 tape-recorded standardized interviews of primary school teachers, I found (Gobo, 1997, pp. 106–107) that respondents attached to each response alternatives:

- five different meanings to ‘strongly satisfying’;
- 15 meanings to ‘fairly satisfying’;
- six meanings to ‘fairly unsatisfying’; and
- five meanings to ‘strongly unsatisfying’.

To describe the most ambiguous response alternative only, ‘fairly satisfying’ was used by respondents to label their work relations in 15 ways or meanings, as:

1. average;
2. partial (i.e. thinking of only a portion of the actor, e.g. the pupils—target of the evaluation);
3. fairness to some pupils;
4. keeping control of themselves in order to disguise their thoughts;
5. constructive and dialectical (meaning positive relations);
6. not conflicting (meaning negative relations);
7. indifferent or neutral;
8. superficial (meaning negative relations);
9. inactive;
10. difficult;
11. conflicting;
12. non-existent;
13. formal;
14. occasional;
15. exclusion (i.e. this response alternative was selected because respondents do not like
the other three response alternatives).

What is impressive about this list of meanings attributed to the same response alternative is that sometimes respondents select this response alternative to state directly opposite statuses (feelings or opinions), thus breaking the scale’s one-dimensionality. In fact, the meanings numbered 6–9 above designate negative evaluations which are opposite to meaning 5 and especially to the response alternative itself which should mean fairly satisfying relations. Furthermore, this phenomenon reveals a response curvilinearity (Coombs, 1953; Edwards & Kenney, 1946).

The multiple word meaning phenomenon is something that survey researchers in contemporary societies can expect to encounter almost everywhere for at least two reasons: the increasing presence of immigrants who are not fully competent in the local languages (Peil, Mitchell, & Rimmer, 1982, Peil, 1983) because they are not (obviously) native speakers, and the great numbers (in some countries such as Italy or France) of native dialect-speaking people.

Pragmatics of Response Alternatives

The cause of these multiple meanings is not only semantic (i.e. the alternative response scales are differently interpreted by respondents) but also pragmatic: respondents use alternative response for particular communicative and practical purposes. As Briggs (1984), Cicourel (1964), and De Santis (1980) have already stated, interviews are dyadic interactions which take place within a social situation embedded in a particular setting. Through the interview, the participants exchange indexical, contextualized, and locally situated meanings, as Suchman and Jordan (1990, p. 241) point out:

The interview is (...) an essentially interactional event. From the moment that the interviewer sits down across from the respondent and begins to talk, the survey interview assumes and relies upon a wealth of conventions and resources from the ordinary conversation.

The interaction transforms an apparently informative event into a communicative activity. Therefore, respondents balance their internal or private statuses (e.g. opinions, feelings, and attitudes) with cultural models (i.e. social conventions, norms, and values) concerned with the realm of public opinion. These cultural models and their social expressions (as diffidence, conspiracy of silence, reticence, demeanour, and the norm of even-handiness; see Table 2) shape the respondent’s selection of the response alternatives.

In the present research, the positive response alternatives (‘strongly satisfying’ and ‘fairly satisfying’) do not express the respondents’ state only but are also used for particular communicative and practical purposes. In the larger sample of 1000 respondents (macro level) a clue to the pragmatics of response alternatives can be found in the statistical distribution of answers to the 10 items concerning personal relations with
<table>
<thead>
<tr>
<th>Effects</th>
<th>Magnitude</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Questions</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Question order</td>
<td>12–15%</td>
<td>Clark and Schober (1992), Kalton, Collins and Brook (1978), and Schuman and Presser (1981)</td>
</tr>
<tr>
<td>Open vs. closed question</td>
<td>11%</td>
<td>Rugg and Cantril (1944)</td>
</tr>
<tr>
<td></td>
<td>13%</td>
<td>Schuman and Presser (1977)</td>
</tr>
<tr>
<td></td>
<td>30–60%</td>
<td>Bradburn and Sudman (1979), Coxon (1986), Schuman and Presser (1979), Schuman and Scott (1987), and Schwarz and Hippler (1987)</td>
</tr>
<tr>
<td>Misunderstanding of question task</td>
<td>?</td>
<td>Gobo (1992, pp. 137–139, 149)</td>
</tr>
<tr>
<td>Wording</td>
<td>Even 30%</td>
<td>Schuman and Presser (1981, pp. 296)</td>
</tr>
<tr>
<td>Items</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item order</td>
<td>?</td>
<td>Schuman and Presser (1981)</td>
</tr>
<tr>
<td>Misunderstanding of items meaning</td>
<td>At least 70%</td>
<td>Belson (1981)</td>
</tr>
<tr>
<td>Object reaction</td>
<td>7–12%</td>
<td>Cacciola and Marradi (1988) and Sapignoli (1992, pp. 125)</td>
</tr>
<tr>
<td><strong>Responses</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Response order</td>
<td>6–14%</td>
<td>Schwarz, Hippler, and Noelle-Neumann (1992)</td>
</tr>
<tr>
<td>Response set (yea-saying)</td>
<td>?</td>
<td>Lentz (1938) and Moun (1988)</td>
</tr>
<tr>
<td>Curvilinearity</td>
<td>?</td>
<td>Coombs (1953), Edwards and Kenney (1946), Fee (1979), and Tourangeau, Rasinski, and D’Andrade (1991)</td>
</tr>
<tr>
<td>Anchoring</td>
<td>?</td>
<td>Ostrom and Upshaw (1968) and Wyer (1974)</td>
</tr>
<tr>
<td>Misunderstanding of response alternative meaning</td>
<td>52%</td>
<td>Razzi (1992, pp. 55)</td>
</tr>
<tr>
<td>Misunderstanding of Likert-type scale tasks</td>
<td>?</td>
<td>Gasperoni and Giovani (1992) and Sapignoli (1992)</td>
</tr>
<tr>
<td>Don’t know (As response alternative)</td>
<td>20%</td>
<td>Schuman and Presser (1981, pp. 120)</td>
</tr>
<tr>
<td>(As response alternative)</td>
<td>47%</td>
<td>Razzi (1992, pp. 48–49)</td>
</tr>
<tr>
<td><strong>Respondents</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Even-handedness norm</td>
<td>5–37%</td>
<td>Schuman and Ludwig (1983)</td>
</tr>
<tr>
<td>Invented opinion</td>
<td>30%</td>
<td>Hartley (1946) and Schuman and Presser (1981)</td>
</tr>
<tr>
<td>Social desirability</td>
<td>?</td>
<td>Edwards (1957) and Kahn and Cannell (1957)</td>
</tr>
<tr>
<td>Extreme response style (Yea-saying/nay-saying)</td>
<td>?</td>
<td>Bachman and O’Malley (1984), Gergen and Back (1966), Hamilton (1968), Harvey (1971), and Rorer (1965),</td>
</tr>
<tr>
<td>Telescoping errors</td>
<td>?</td>
<td>Sudman and Bradburn (1973)</td>
</tr>
<tr>
<td>(Vote misreport)</td>
<td>55%</td>
<td>Silver et al. (1986)</td>
</tr>
<tr>
<td>Forgetting of voting</td>
<td>5%</td>
<td>Brehm (1993, pp. 16)</td>
</tr>
<tr>
<td>Interview refusal (academic research)</td>
<td>20–33%</td>
<td>Brehm (1993, pp. 16)</td>
</tr>
<tr>
<td>Interview refusal (market researches)</td>
<td>30–50%</td>
<td>Crespi (1988)</td>
</tr>
</tbody>
</table>
Most of the respondents (80–90%) selected positive response alternative (‘strongly satisfying’ or ‘fairly satisfying’) for each item. However, taking into account the strikes occurring during the 1980s and 1990s, the high percentage of temporary jobs in the Italian school system, and the low prestige attributed to the social category of teacher by the respondents themselves in a question contained in the questionnaire,8 Italian teachers appear to be one of the most dissatisfied occupational categories.

Listening to the 80 tape-recorded standardized interviews (micro level), I found several inconsistencies between respondents’ selection of response alternative (generally positive) and their verbal comments (generally negative) which preceded or followed the selection, as shown in the following excerpt:

INT Your relations with your colleagues in the teaching body are strongly satisfying, fairly satisfying, fairly dissatisfying, or strongly dissatisfying or nonexistent?
R Er, well the teaching body, okay it’s not that I don’t give my opinion but well, uh … I’m accommodating really. Perhaps what most people say and I think so too, is that I’m a bit of a rebel … but er … anything for a quiet life.
INT So, what are they like, that is in your opinion?
Rh (a little reluctant but smiling)
INT No, in your personal opinion what are they like? (reading again the four response alternatives)
R Oh okay … strongly dissatisfying … strongly satisfying for heaven’s sake.
INT (smiling) no, but just what you think, in any case it’s not that … (the interviewer, understanding the respondent’s diffidence, tries to reassure her that the answer will be kept secret)
R Mmm (declining the interviewer’s move) anyway …
INT strongly, fairly satisfying …
R no, strongly satisfying.

Table 2. (Continued)

<table>
<thead>
<tr>
<th>Effects</th>
<th>Magnitude</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question refusals (missing) in:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(a) question about vote</td>
<td>30–40%</td>
<td>Calvi and Vannucci (1995)</td>
</tr>
<tr>
<td>(b) question about monthly salary</td>
<td>58%</td>
<td>IARD (1993, p. 249)</td>
</tr>
<tr>
<td>Synchronic response inconsistency (in the same interview)</td>
<td>19–48%</td>
<td>Gasperoni (1993, pp. 13–14)</td>
</tr>
<tr>
<td>Diachronic response inconsistency (panel)</td>
<td>17–82% due to salience</td>
<td>Hochstim and Renne (1971)</td>
</tr>
<tr>
<td>Mood of the day</td>
<td>37%</td>
<td>Gobo (1992, pp. 263)</td>
</tr>
<tr>
<td></td>
<td>?</td>
<td>Abbey and Andrews (1985)</td>
</tr>
</tbody>
</table>

*aThe question mark means several things:
• the magnitude has not been counted yet;
• there is not a general tendency;
• the measurement of that particular phenomenon suffers from numerous methodological problems;
• the phenomenon has been discovered in seminal methodological research only;
• it is based on methodological ethnographic accounts only.
Respondents can show deference or compliance to the interviewer because the latter can be perceived as a person who can make a change (even if only minor) in the respondent’s life (Atkin & Chaffee, 1972; Boccuzzi, 1985, p. 246; Collins, 1970, p. 442; Hyman, Cobb, Feldman, Hart, & Stember, 1942, p. 246; Kahn & Cannell, 1957; Pitrone, 1984, p. 106).

The use of response alternatives for communicative purposes can be caused by the 'yea-saying' style. This phenomenon leads to different response behaviours: the response set and the extreme response style (Gergen & Back, 1966, Hamilton, 1968; see Table 2). As many authors have stated, Likert-type scales are particularly subject to the acquiescent set because they have the same response alternatives for long lists of items.

Finally, another use of response alternatives for communicative purposes is represented by the 'norm of even-handedness' (Schuman & Ludwig, 1983; see Table 2).

Three Remedies

Even if it seems difficult to reduce pragmatic response effects, the semantic part of the multiple-word-meanings phenomenon related to response alternatives can be minimized.

The first remedy has already been proposed above, that is to adopt a ruler on which to locate the response alternatives (Amisano & Rinaldi, 1988; Blackburn & Mann, 1975). The ruler forces the respondents to perceive as equal the distance among response alternatives. This is a crucial pre-requisite in order to correctly exploit the cardinal properties of numbers used as codes (values) of response alternatives, especially when we create an index. On this topic, Marradi (1995, pp. 77–78) suggests embedding in the numerical codes (attached to response alternatives) our knowledge and evaluation about the distance which lies between the response alternatives, as in the following examples:

<table>
<thead>
<tr>
<th>Frequency of church attendance</th>
<th>Degree of liking of the personage X</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 Never</td>
<td>0 Not liked at all</td>
</tr>
<tr>
<td>2 Christmas and Easter</td>
<td>2 Not strongly liked</td>
</tr>
<tr>
<td>3 Once at month</td>
<td>3 Mildly liked</td>
</tr>
<tr>
<td>4 Every Sunday</td>
<td>4 Strongly liked</td>
</tr>
<tr>
<td>7 Everyday</td>
<td>6 Very strongly liked</td>
</tr>
</tbody>
</table>

Relatively to the frequency of church attendance, the one who goes everyday has a quite different attitude in comparison with the one who goes on Sundays only; in the same way, the one who goes at Christmas and Easter only has a very different attitude as regards the one who never goes. The distance among the middle categories can be evaluated as smaller as in the example of liking a personage where the threshold between ‘not very liked’ and ‘not liked at all’ can be perceived as bigger than the others, as well as the threshold between ‘mildly liked’ and ‘very strongly liked’.

Obviously, someone can criticize these evaluations and propose other codes; or they can attach—as in the traditional practice—the series of natural numbers (1, 2, 3, …)
avoiding every sophistication. However, as Edward Tufte (1970, pp. 440–441) underlines, attaching to response alternatives the series of natural numbers as codes is as arbitrary as every other decision that fixes a monotonic relation between the series of response alternatives and the series of numerical codes attributed to them. This is true whether we want later to apply data-analysis techniques suitable for cardinal variables or use techniques appropriate for ordered variables only. It is preferable to attribute the series of cardinal numbers only if we are reasonably positive that the distance among the response categories is approximately equal.

These considerations must be extended also to the weight of negative response alternatives. In data-analysis practice, positive response categories are considered comparable to negative response categories. However, in terms of social relations, selecting a positive answer does not have the same effect as uttering a negative answer: it is easier to agree than to disagree (Bishop et al., 1982; Campbell, 1975; Schuman, 1966). As a matter of fact in the standardized interviews, highly educated respondents, because they are the only ones who tend to use the whole range of response alternatives, are inclined to use negative categories. In a survey, as in everyday life, selecting a negative response alternative is interactionally much more embarrassing than selecting a positive response alternative. In the latter case, the selection is often followed by ‘remedial interchange’ (Goffman, 1971), a respondent’s verbal comments justifying themself as having manifested a socially undesirable opinion or having departed from the ritual of discretion. This interactional phenomenon also explains the reason why, in graphs 1 and 2, the three negative response alternatives are perceived as very close, and in contrast respondents perceive a greater distance between positive response categories. In other words, ‘there is just one “no” and many types of “yes”’. For these reasons, negative responses cannot be considered symmetrical (or equidistant) to positive responses, and negative categories need codes capable of embodying the social conventions of the situation wherein the standardized interview is embedded.

The second remedy, complementary to the previous remedy, is more difficult (but not impossible) to pursue and requires prior ethno-linguistic research to substitute traditionally worded response alternatives with terms taken from everyday-life conversations and used by native respondents to label and classify statuses of agreement/disagreement, like/dislike, satisfying/dissatisfying, etc. This strategy aims to reduce the cultural gap between researchers and sub-sets of respondents about the meanings attributed to response alternatives, and respondents’ differing ways of articulating and expressing attitudes, opinions, and feelings. The issue is crucial also in understanding and minimising the phenomenon of ‘response inconsistency’ in panels.

Traditionally, the causes of inconsistency between administrations have been explained by:

1. natural change in the respondent’s attitude in the course of time (Converse, 1970);
2. induced change, that is the change has been affected by the previous interview because the respondent has had time to reflect on the questions (Bohrnstedt, 1970, p. 86; Campbell & Stanley, 1963, pp. 20–22; Carmines and Zeller, 1979, p. 39);
3. respondent’s mood on the day (Abbey & Andrews, 1985; Atkinson, 1982; Moun, 1988; Rodgers & Converse, 1976), anxiety (Lewis & Taylor, 1955), boredom and apathy (Hyman et al., 1942; Pinto, 1964, p. 674);
4. previous experience (good or bad), so the respondent’s reaction in the second administration can change (Scott, 1968);
5. forgetting information, due to the time interval between administration;
6. different interviewer’s behaviour during the two administrations;
7. respondent’s unreliability because they do not have ‘well-formed attitudes’ (Converse, 1970) or ‘well-structured schemas’ (D’Andrade, 1989). This kind of actor has been called a ‘mass respondent’ (Rose, 1950) or ‘floater’ (Schuman & Presser, 1981, p. 146). However, what has been neglected is that part of such an inconsistency could be caused by the questionnaire format itself, particularly:
8. multiple word meaning of items which can be interpreted differently by the same respondent over the course of time;
9. multiple word meaning of response alternatives;12
10. semantic interchangeability of response alternatives, as perceived by subsets of respondents. Where response alternatives of Likert-type scales are not commonly used in ordinary language by respondents, their meanings are so shaky that respondents can select categories without consideration, using different response alternatives to label the same feeling or opinion. In re-interviewing primary school teachers in Naples, I (Gobo, 1992, p. 255) found that notwithstanding the fact that the respondent’s comments were similar in the two administrations, they selected different response alternatives several times because these (apparently standard) categories (such as ‘fairly satisfying’) were linguistically and cognitively distant from their language and ways of thinking. Here, are three further examples:

Boccuzzi (1985, p. 251) found that in Taranto district (Italy), many respondents were puzzled about using ‘satisfying/dissatisfying’ response alternatives to label their feelings about the job: ‘it was clear they were not accustomed to associate with work an adjective referring to pleasure’. Generally speaking, low-educated respondents are not able to use Likert-type scales, that is to align their thoughts to response categories like degree of ‘consent’ or ‘favourable’. They conceptualized issues using ‘true/false’ or ‘right/wrong’ categories (Pitrone, 1995, p. 55).

Interviewing 100 parents in Lombardy (Italy), Lanzetti (1993, pp. 28–29) notes: ‘only 23% of respondents correctly used the response alternatives without problems; 30% could not make it out four times out of ten; 47% (mainly less educated respondents) had problems every time.’ These ethnographic accounts highlight the need to use words taken from respondent’s ordinary language as response alternatives.

The third remedy is to broaden the interviewer’s tasks. The role of interviewer has long been discussed in the survey literature. As J. M. Converse (1987, p. 95) reminds us, in the 1920s and 1930s some academic and especially a good deal of the market research literature ‘placed the interviewer in some sort of middle ground of freedom and responsibility, with questions less standardized (…) There was concern that trying
to standardize the interview more fully might interfere with the communication process.’

The interviewer was advised to act responsibly, with freedom of ‘conversationalizing’ questions without modifying their meaning. The directors of market research studies believed that the standardization of interviewer’s behaviour was mandatory in laboratory experiments but that it could not work in interview situations where incessant adaptation of the questionnaire to respondents and social situations is necessary. At the beginning of the 1950s, this wise practice was replaced by another practice (standardization) affected by the behaviourist perspective (Hamilton, 1929; Rice, 1929), and it still affects contemporary survey methodology. According to this practice, the response alternatives must be selected by respondents only. However, respondents are often biased in their interpretation of the meaning of response alternatives which are often quite different from the meanings attributed by researchers. Because the assumption about a natural correspondence between psychological and numerical intervals has not yet been proved (Pawson, 1982: p. 54), and formal languages are incompatible with natural languages (Pawson, 1982, 1983), the third remedy proposed here is to let interviewers act (when respondents have difficulty in using scales or formatted answers) as interpreters in order to make them responsible for selection on behalf of the respondent (Galtung, 1967, p. 120) and for (always imperfect) translation from the respondent’s ordinary language to the formal or mathematical language underlying measurement scales: ‘Interviewers should be trained in the concepts inherent in the questions and be allowed to probe, rephrase, and adapt the questionnaire to individual respondent needs’ (Groves, 1989, p. 404).

Interviewers and respondents should work together to ‘jointly construct’ the meaning of questions and answers (Mishler, 1986), because there is a conflict between interviews as conversation and interviews as data collection, due to the fact that surveys (in order to succeed) rely on conversational norms which suppress ‘interactional resources that routinely mediate uncertainties of relevance and interpretation’ (Suchman & Jordan, 1990, p. 241) in conversations. As Schober and Conrad (1997) have shown, in their laboratory experiment with trained telephone interviewers using either standardized techniques and flexible interviewing, there is no substantial difference in response accuracy when the concepts in the questions are clearly mapped onto the fictional situations of respondents. In addition, and even more interesting, when the mapping was less clear, flexible interviewing increased accuracy by almost 60%.

However, opposition to this remedy is widespread in the literature, even if it seems based more on a methodological narrative than on data. Clausen (1968), Feldman et al. (1951), Hauck and Steinkamp (1964), and Hyman et al. (1942) re-appraised the alarmism about the effect of interviewer’s attitudes, opinion and personality on respondent’s answers and showed a bias ‘of moderate magnitude’ (Hyman, Cobb, Feldman, Hart, & Stember, p. 244), that is about 10%. As regards Hyman et al.’s theory about interviewer’s ‘attitude-structure expectations’ (1942, p. 59), ‘role expectations’ (1942, p. 62), and ‘probability expectations’ (1942, p. 64), Hagenaars and Heinen (1982, p. 125) write that Hyman’s statements are based upon few data. In the same way,
G. Gobo

Sudman and Bradburn (1974, p. 138) believe that the biases introduced by questionnaire tasks, social desirability, forgetting, and so on, are more dangerous than interviewer’s behaviour. Bradburn and Sudman (1979, pp. 50 and 171–172) conclude that interviewer’s errors do not have remarkable effects on the quality of data. Dohrenwend and Richardson (1956) argue that most errors are caused by overly tight control and that interviewers must learn to be more responsive to respondents. Penneff (1988) provides a very revealing insight into practice on the ground. By observing interviewers involved in a large field survey, Penneff maintains that in reality, survey interviewers adopt the qualitative interviewer’s skills. They try to interest respondents by letting their own personalities show, interact in a non-neutral way, and add personal comments to avoid misunderstanding or refusal. This should not be seen as cheating but as adapting the interview process to the subject’s definition of the situation. The interviewers studied by Penneff, who were regarded by the field survey director as his most successful interviewers, proved to be those who used these techniques the most (the survey director was appalled when Penneff told him this). Penneff uses this finding to argue that, in practice, the methods of qualitative sociology penetrate the survey interview.

Nevertheless, the dogma about standardization is still alive. In a survey conducted in a valley area of South Wales, Michael Brenner—following the method of Cannell, Lawson, and Hauser (1975)—has documented that approximately 30% of questions were not correctly asked by interviewers, and this bias had a negative effect on 13% of responses (Brenner, 1982, p. 155). However, this result is inconsistent with the data presented by Brenner later: ‘when these questions were asked directly so that definite answers were suggested to the respondents’, the percentage of answers considered adequate (by the author) increased to 20% (Brenner, 1982, p. 157). This finding means that the interviewer’s directive style produces a remarkable increase in the response quality, larger than the methodologically correct non-directive style of asking questions. The complementary results of Brenner’s research indirectly document the thesis of this article: it is not the interviewer’s non-standardized behaviour itself that is responsible for response errors, but only some incorrect moves of the interviewer in modifying question meaning, fast reading of questions, omitting to use the card required with the question, and so on.

Dijkstra and van der Zouwen (1988) have replicated both Cannell et al.’s (1975) and Brenner’s (1982) studies. In relation to interviewers, they found: a mean of 4% of deviations from the questionnaire; 8% of irrelevant behaviours; a mean of 10% of hinting from the questionnaire (p. 30) (and if the interviewer has to probe further, the percentage of suggestive questions posed by the interviewer ranges from 15 to 23%) (p. 31); and finally 16% of choosing behaviours (on behalf of respondents). Because an interviewer’s error does not necessarily produce a response error, the crucial question is: how much do the interviewer’s errors really affect the data quality? In other words, as Schober and Conrad (2002, p. 69) underline:

Since interviewers always influence responses, this raises the question of which kinds of influence are benign and which are not. We argue that the criterion should be how interviewer behaviours affect response accuracy—that is, how well responses correspond with the definition the survey author had in the mind.
For this purpose, it is important to recall Beatty’s (1995, p. 154) statement:

We are, after all, interested in reducing total error in surveys. If attacking the slightest interviewers deviation brings about modest reduction of interviewer error—but simultaneously causes a great increase in error from the respondent, who is unable to draw on the communicative resources of an informed, intelligent interviewer—then the strategy is self-defeating.

Comparing the performance of bad interviewers who made many errors with the style of good interviewers who made few errors, Dijkstra and van der Zouwen’s (1988, p. 32) results show that the size of bias introduced by bad interviewers is not so remarkable as to point to the interviewer’s performance. As a matter of fact, even a serious error such as ‘choosing behaviour on the part of the interviewer appeared to have the least effect in the observed relations between respondent type and respondent answer’.

Directing too much attention to interviewer’s effects is, to use an old proverb, like seeing the tree and missing the forest. The forest is the long list of biases mainly imputable to the questionnaire (or to the researchers as its designers) and to respondents (see Table 2).

**Conclusion**

In the light of these data, we can conclude that the interviewer’s errors are of secondary importance and far smaller than the researcher’s and respondent’s errors. As Bradburn (1983, p. 291) states: ‘the characteristics of the task(s of the questionnaire) are the major source of response effects and are, in general, much larger than effects due to interviewer or respondent characteristics’.

In addition, Schaeffer (1995, p. 83) reminds us:

criticisms of traditional standardized interviewing are particularly effective when taken together with research which suggests that the recall of events may be improved by procedures that do not fit neatly within the linear structure of standardized interview (Means, Swan, Jobe, & Esposito, 1992), that a less formal style of standardized interviewing may be more motivating (e.g. Dijkstra, 1987), and that interviewers do not always implement standardization well (…) and a formal standardized interview may not be the best social environment for stimulating and motivating recall of complex topics

This consideration leads to the problem of how to improve data quality. Rather than pursing the illusory goal of improving the wording only in order to reduce the necessity of interviewer’s probes (as suggested by Fowler & Mangione, 1990, p. 46) because ‘total elimination of interviewer error is impossible’ (Beatty, 1995, p. 155), we can achieve data quality by entrusting to the interviewer a more active role in order to bridge questionnaire and respondent, and to reduce the gap between researchers’ and respondents’ meanings. In David Riesman’s (1958, p. 305) words: ‘the task of the interviewer, as I see it, (is) to adapt the standard questionnaire to the unstandardized respondents’.

Several studies have evidenced that standardizing the stimuli (i.e. questions, items, response alternatives, and interviewer’s behaviour) does not necessary imply the standardization of their meanings which should remain the main aim of every data collection. As Houtkoop-Seenstra (2000, pp. 180 and 182) concludes:
Having studied tape-recorded standardized survey interviews for some years now, I have become increasingly convinced that the quest for standardization is no longer tenable, at least if its purpose is to generate not only reliable, but also valid, research data (…) We should allow interviewers-as-spokesperson to discuss with respondents the intended meaning and purpose of questions, as well as the respondent’s answers. This discussion may increase the validity of the research data, even though a more flexible way of interviewing may at times cause inappropriate interviewer behaviour, such as presenting the respondent with leading questions.

Clearly, the aim of reducing respondent’s errors by broadening interviewer’s tasks will lead to an increase in interviewer effects. However, the dilemma is which kind of errors we prefer to minimize. In addition, the magnitude of interviewer’s errors is far smaller than those of respondent’s because:

- a trained interviewer knows the purpose and correct meaning of questions, items, and response alternatives better than respondents;
- the meanings in the (relatively small) ‘interviewers’ community’ are more consistent than in the mass of socially and culturally different respondents.

In other words, the questionnaire and interviewer’s behaviour must be user-centred and really tailored to respondents and their differences. Does standardization help to succeed in this aim?

Notes

[1] This paper has been presented at the ISA (International Sociological Association) XIV World Congress, Research Committee 33 ‘Logic and Methodology, Montreal, 31 July, 1998, and at the ISA Fifth International Conference on Social Science Methodology, Cologne 3–6 October, 2000. I want to thank Aaron Cicourel, Marco Razzi, and four anonymous referees for helpful comments and suggestions.

[2] In statistics, there are two variances: between and within groups. Transposing these two concepts, we can say that cognitive problems are analogous to the variance within groups (WSS) because they are cross-sample. On the other hand, social problems appear differently among social categories of people, and it is useful to explain the variance (BSS) with sociological, psychological, or demographic variables.

[3] The same cognitive/semantic phenomenon has been noticed by Blasius and Thiessen (1998).


[5] They are a section of a national, probabilistic, and representative sample (1000 cases) of elementary-school teachers. The teacher study was done as a piece of substance research. However, this category has a special methodological interest due to the fact that it is a fairly highly educated social category. In other words, if we find that educated people have trouble with response alternative, what about people who do not have a degree? The method is based on a discourse analysis conducted on the audio-recorded materials: the interviews have been transcribed, focusing particularly on the comments which follow each answers. When, during the interview, there was not a direct comment, the interviewers asked: ‘what do you mean as (the response alternative)?’ pronouncing the category selected by the respondents. Then, the author has accurately classified the meanings.

[6] The response is a balance among the relations with all pupils.
[7] The concern of the six respondents when they select this response is to not give an overly severe judgment which could be an outcome unfair to the pupils with whom the respondents have excellent relations.

[8] The question was: ‘In your opinion, in the next 10 years, will the prestige (i.e. the social consideration generally speaking held by teachers) increase, remain the same, or decrease?’ The last response alternative was selected by 61% of respondents.


[12] This difficulty in solving bias is well known in comparative cross-cultural surveys (see Bulmer & Warwick, 1983; Peil et al., 1982).

[13] It is beyond the purpose of this article to stress the serious interactional problems which a standardized behaviour poses. Interviewers who strictly follow the rules of standardized interviewing (as stated e.g. in Bailey, 1978; Brenner, 1985, p. 19; Fowler, 1984; Fowler & Mangione, 1990; Survey Research Center of Berkeley, 1990) frequently present themselves to respondents as impolite, insensitive, and unintelligent individuals because they ask redundant questions. Houtkoop-Seenstra (2000, p. 183) proposes: ‘we should give interviewers the freedom to draw inferences and then verify them with the respondents. If a respondent [in a previous comment] mentions ‘my husband’, the interviewer should not ask whether the respondent is ‘single, married, a widow, or living together’ but should be allowed to verify that the respondent is married (...) in a leading manner’.

[14] Houtkoop-Seenstra (2000, p. 182) suggests allowing interviewers to accept unformatted answers because respondents ‘have a hard time remembering the list of response options for the duration of the interview. When interviewers respond to an unformatted answer by re-offering the response options, thus implicitly informing the respondent how he or she should answer the questions, the transcripts show that respondents follow this rule for only a short time. A few questions later we find them providing unformatted answers again’.


References


Hörmann, H. (1983). The calculating listener or how many are einige, mehrere, and ein paar (some, several, and a few)? In R. Bauerle, C. Schwarze & A. Von Stechow (Eds.), Meaning, use, and interpretation of language (pp. 221–234). Berlin: de Gruyter.


