Journal of Official Statistics
Vol. 1, No. 2, 1985, pp. 159–178

Testing/Assessing Question Quality – Some Swedish Experiences

Mats Thorslund[1] *and Bo Wärneryd*[2]

Abstract: The article describes some methodological studies related to survey question quality conducted at the Survey Research Institute, Statistics Sweden. Approaches to studying question quality vary in the degree to which they try to understand the survey process or to measure quantititative aspects of it. Methods discussed include replicated measurements, systematic question scrutiny, and validation studies.

The effects of factors confounding results in different studies are evaluated. Examples in the text concern mainly large continuing surveys, where special opportunities for quality assessment often occur. The discussion of efficiency aspects of different methods is applicable, however, even in smaller and/or unique surveys.

Key words: Survey question testing/assessment; reliability; validity; efficiency.

1. Introduction

1.1. Background

In this article, some methodological studies related to survey question quality will be discussed. These studies have been conducted at the Survey Research Institute, Statistics Sweden. The Institute is responsible for several large surveys, among them the Survey of Living Conditions.

When testing/assessing question quality is concerned, the newly published handbook on questionnaire development (DeMaio (1983)) is a natural starting point. This handbook summarizes much of the repertoire from which we have drawn. The summary reflects the methodological impact of many large-scale surveys. We see, however, some shifts in emphasis during the last 20 years or so. To some extent, we believe that this shift depends on the increasing number of continuing surveys (but not necessarily panels), and would discribe it as a shift from meeting needs of statistical information by large-scale one time surveys.

At the Survey Research Institute, the continuing data collection has focused on the need to assess question quality. The traditional way of making a one-time survey involved a more or less ambitious pilot survey with, when economic resources and time allowed it, unstructured individual interviewing to develop ideas of survey questions, informal testing of early drafts, realistic pilot studies with tape recordings, and interviewer debriefings as outlined in the handbook. In the early seventies, the repertoire was broadened to include variants of

[1] Department of Social Medicine, University of Uppsala, Sweden.
[2] Survey Research Institute for Statistics on Living Conditions, Statistics Sweden.

Belson's methods of following up respondent's interpretation of survey questions (see e.g. Belson (1968)). Somewhat later systematic observations of interviews, reinterview studies, and procedures for scrutinizing questions have been used. It is these efforts as well as a special validation study that will be discussed in this article.

Despite the handbook and a growing literature in this field, there seems to have been very little systematic study made of the efficiency of different methods of testing/assessing quality. We have seen only one small study where deliberately "loaded" questions have been used to test different methods' (Hunt et al. (1982)) efficiency in detecting errors.

A theme for this article is that the traditional methods are lacking in efficiency – at least, in the way they have been used in developing the Swedish Survey of Living Conditions (discussed in Thorslund and Wärneryd (1985)). Efficiency may be defined as the capability to distinguish between random errors and systematic errors, or, to use terms associated with measurement, lack of reliability and validity. Some aspects of the efforts/costs involved will be discussed.

Two main aspects of question problems – issue definition and language – will also be used in discussing the efficiency of different methods.

We will make no strict distinction between studies dealing with the testing of question quality in developing questionnaires and the assessing of quality of survey questions already in use.

1.2. The Survey of Living Conditions and the Measurement of Change

Our discussion will frequently refer to the Swedish Survey of Living Conditions (SLC). The SLC is based on nation-wide samples of the population in the ages 16–84 (1974–1979, 16–74). A national sample of 8 000–12 000 persons are interviewed each year. Altogether the survey comprises about 500 indicators. The interview data are supplemented by data on, for instance, income, money transfers and taxes from various registers (for an overview, see Vogel (1981, 1982)).

In principle, the welfare indicators sought through the SLC are objective measures, that is they should not be coloured by the respondents' ambitions and reference frames. This means that items such as wishes, demands, and opinions are not formally surveyed. In this respect, the Swedish surveys differ from other similar projects.

The questions and variables in these surveys have to a large extent been used before, in surveys covering special topics and in the first Swedish level-of-living survey from 1968. For example, the questions on health care utilization are to some extent based on the American Health Interview Surveys (see e.g., Cannell et al. (1977)) and the British General Household Survey (The General Household Survey (1973)). Within other areas, for instance, measurement of the actual ("objective") work environment, there has been no obvious source to borrow tested variables from.

During the establishment of the Swedish Surveys of Living Conditions, the methodological work has mainly been oriented towards the development of well-functioning interview questions and towards a well-controlled system of data collection. Problems of nonresponse have received a great deal of attention.

As well-functioning questionnaires were developed and as nonresponse was kept at an acceptable level, the methodological work became focused on measuring and improving the reliability and validity of the variables included. This "new" orientation can also be seen as a consequence of heightened demands put on the data for use in comparisons over

time. On the one hand, randomness in response on health status, (at one measurement, a certain proportion of respondents are categorized as "suffering from long-term illness," a few weeks later the same respondents are categorized as "healthy" – and vice versa) often means that the errors cancel each other out, so that the total estimates from the two measurements do not differ (Hochstim and Renne (1971)). On the other hand, this same type of randomness may be disastrous if we want to follow individuals over time to describe patterns of "recruitment" into the illness category.

Comparisons over time entail other problems of analysis and presentation. If, a difference between two years has been statistically established, one must try to explain why this difference was obtained. Can it be a consequence of the interviewers' having changed their handling of the question, that the sequence of questions in the questionnaire has been changed so that respondents perceive the question differently, or that coding practices have changed? All these types of changes may influence the results, and one must strive to keep them at a minimum.

On the other hand, every survey is part of a cultural setting where norms may change. Suppose that the percentage of people in the labour force who regard some aspect of their working life as "bad" has increased (statistically significantly) from, say 1975 to 1984. It is not possible to explain this as an artefact due to changes in measurement procedures. So, is there an actual change in work conditions in this respect. Or has an increased awareness – people in general have raised their criteria for what a workplace should be – caused the difference? Or a combination of these factors? Or can the cause be found in structural changes in the labour market which have taken (and is still taking) place: that proportionately more workers are employed at large places of work, within certain branches of the economy, etc.

Studies of change over time thus contain many methodological problems. A number of studies dealing with aspects of measurement of change were initiated at Statistics, Sweden in 1978[3]. Some of them will be presented in Sections 2.3 and 3.

1.3. On Measurement and Understanding

There are two basic approaches to testing/assessing question quality.

i) measurement of some quantity reflecting amount of 'error' at some point in the data collection; and
ii) study of the process itself, to try to understand what aspects of the process that lead to errors.

To some extent, these approaches are irreconcilable in practice. To study processes is expensive, cumbersome and quantification is hard. Process study often means case studies or at least very limited statistical counting. Measurement means larger samples, studied using less expensive methods at check points in the process, in other words at points where things can be counted. Christoffersen (1984) compared quantitative methods of comparison with a "black box model." Here you know what is put in – variations in data collection methods – and what the output is – distributions of survey answers. But the process producing answers is contained in the "black box." This may be a somewhat unjust description of quantitative studies of data collection, since, as will be seen later, even quantitative studies may contain data from inside the "black box."

However, Christoffersen, as we do, recommends the use of both approaches in any given survey problem. In a standardized survey,

[3]Reports on these studies are in Swedish and will not be referred to here.

there is little insight in the question-answer process, unless special steps are taken. To follow how answers are generated means following processes, and that, as already noted, is hard to quantify as well as being costly. The main possibilities are either to try ex post facto to reconstruct what happended or to be there when it happens and "take notes." Examples of the first kind are different kinds of follow-up interviews, where questions are asked about how questions in an earlier interview were perceived and how answers were determined. Belson is usually referred to in this context for his basic study of the quality of a readership survey (Belson (1963)). The techniques developed by Belson (see Belson (1968 and for examples of results, also 1981)) have been adopted in some studies at Statistics Sweden.

However, "frame-of-reference probing," as it is called by the handbook on approaches to developing questionnaires, has its limitations in relation to the direct study of the process. In interview surveys, tape recordings can be used to study the process. This has mostly been done in pilot surveys. No special problems have been met in getting respondents to accept the recorder in personal interviews. Tape recordings is a relatively cheap method of collecting data on question quality aspects. But it is costly to get the information out of the recordings. Still, parts of what happens in an interview, nonverbal interaction between interviewer and respondent – are not covered. An observer who can take notes of what happens in the interview is needed.

Interviews have been observed in the SLC since 1974, initially to evaluate the practical problems of using an observer in the interview situation and to refine the technique of observing. Thus, it became possible to systematically study how the interview questions were put, perceived, and answered, and what factors caused problems.

In sum, these observations have shown many examples of what can go wrong in interviews, though they may be well planned and tested in pilot studies. Of course, the observations are not well suited to quantitative estimates of "misbehaviours" in the interviews. Rather, they have a sensitizing effect. Increasing our understanding of what takes place in the interview situation is their greatest value.

2. Replicated Measurements

2.1. Uses

Replicated measurement has a long tradition in psychology in the form of test-retest for quantitative evaluation of the reliability of tests. Application to pretesting of survey questions was suggested by Sletto (1940). Its basis is independent measurements at different but equivalent occasions.

Reinterviews for testing/assessing question quality have been used in several surveys at Statistics Sweden. Reinterview studies in the Swedish Survey of Living Conditions and in the Labour Force Survey will be discussed. In our opinion, the results from these studies have been valuable in showing the amount of shakiness in different procedures and by showing that there are considerable variations in inconsistency between survey questions. However, the usefulness of reinterviews depends on the extent to which some basic assumptions are met.

2.2. Basic Assumptions and Do They Hold in Practice?

We want reliable data in survey research. Test theory gives us a measure of reliability as the test-retest correlation, given certain assumptions (Bohrnstedt (1983)). That is, assumptions of "uncorrelated errors of measurement," and so on. There are a number of difficulties tied to these assumptions, when survey data are concerned (see discussion in Bohrnstedt (1983) and Rodgers et al. (1982)).

Replicated measurement means using the same measurement procedure or measuring the same variables at least twice on the same subjects. Furthermore, the measurements should be independent. The prerequisite of independence leads to aiming for long intervals between measurements. Also, it is generally found that the shorter the time interval, the higher the stability found. If the interval is short, the respondent may remember his earlier responses, thereby making them appear more consistent. With a long interval, problem arise when questions are tied to reference periods, which respondents may find hard to remember. Also, the probability of actual change between measurements increases. When using a short interval, procedures leading to independence may include changing context or question order or question form – making the respondent "naive" again – but this must not clash with the prerequisite of measuring the same variable. Thus, it is an intricate problem to find a proper balance between competing demands.

There are further confounding factors in practice, which make it difficult to etablish that the same variables have been measured. The problems we have met will be discussed later. We think one more aspect is worth noting. We cannot disregard the variations in nature of phenomena asked about. We require that certain questions give identical results, regardless of the measuring methods used.

2.3. The Use of Reinterviews in the Survey of Living Conditions

2.3.1. General outline of studies

Since 1979, four reinterview studies have been conducted. They have been made according to the same principles: a sample of those persons interviewed in the regular SLC were reinterviewed about 3 weeks later about parts of the earlier interview. About one-fourth of the questions were chosen from the SLC questionnaire.

Reinterviews were made by telephone. For one part of the 1979 sample, the replication was done by mail survey. Sample sizes have been around 500. Nonresponse was highest in the first study (11%). In the later studies, it has been around 4–5%. The samples do not include those interviewed in the SLC follow-up telephone phase (mainly those who refused a personal interview).

Reinterviews were done by interviewers regularly working in the telephone phase. Generally, it can be maintained that the re-interviews were simpler than regular SLC interviews by being shorter and by the fact that the technically hardest questions were excluded, as were potentially sensitive questions.

Ordinarily, you do not try to take account of actual changes between measurements in test-retest studies. But the SLC questions vary widely in susceptibility to change and the interval between measurements would ideally be variable to account for this. In the reinterview studies, special questions were asked. These questions were of the form "Has there been any change in your housing conditions/family/employment/health/smoking habits, etc., since the previous interview?" Of course some questions have dealt with issues where no change would be expected. In the analysis, people stating that conditions have changed have been excluded.

2.3.2. Some results

Overall, there were more significant differences in response distributions than expected by chance. This especially holds for the 1981 study. However, it is hard to find any pattern in the shifts in marginals. A consistent fact is that the question on long-term illness, included in both 1979 and 1983, yields fewer cases in the reinterview.

Table 1. Consistency Between Regular Interviews and Reinterviews in the Survey of Living Conditions

Question/variable	g	M	Net change
Disposal of dwelling (ownership, tenant ownership/tenancy	2	.04	2
Do you smoke daily? (yes/no)	2	.05	–
Socio-economic status:			
Own (worker/otherwise occupied)	3	.06	–
Father's (worker/otherwise occupied)	12	.24	–2
Is there any place adjoining the house where the people who live there can sit down, e.g., to sunbathe or have a cup of coffee? (yes/no)	17	.39	–1
Does your job involve many repetitive and monotonous tasks? (yes/no)	21	.43	3
Have you been the victim of an accident at work on any occasion during the last 12 months? (yes/no)	7	.52	–
Is it usual in your neighborhood that people exchange a few words when they meet? (yes, very usual/yes, rather usual; no, rather unusual; no, very unusual; don't know)	29	.71	3

In Table 1, some results from the three first reinterview studies are given.

Measures relate to dichotomized data, i.e. four-fold tables resulting from cross tabulations of answers from the original interview with those in the reinterview:

		Reinterview 0	Reinterview 1	
Original interview	0	a	b	$a+b$
	1	c	d	$c+d$
		$a+c$	$b+d$	n

Measures used are:

Gross difference rate, $g = \frac{b+c}{n} \cdot 100$,

Index of inconsistency,

$$M = (b+c) / \left\{ \frac{(a+b)\ (b+d)}{n} + \frac{(c+d)\ (a+c)}{n} \right\},$$

and

Net change, $\frac{b-c}{n} \cdot 100$

In M, the actual gross difference is related to the expected one, given the marginals and assuming independence.

M can be regarded as a measure of association in the table. If we form $1-M$, we obtain an association measure ranging from 0 to 1 and coinciding with the product-moment correlation and the Phi-coefficient for the special case of identical marginal distributions. In this case, M is also identical to the index of inconsistency used in the U.S. Current Population Survey.

The quantities g and M are used as indications of question quality. If they are large, this leads us to look more closely at question content and form, to do observations, etc. Large, unsystematic deviations may be canceled out when the distribution of only one variable is studied, but they become systematic in their effects when variables are crossed and correlations studied. Lack of reliability in a background variable may mean that differences between, say, socio-economic categories in some dependent variable are underestimated.

The useful information from reinterviews are not limited to the quantitative aspect of response deviations. It is astonishing how much you learn from careful and simultaneous examination of answers. Since it is important in our case to eliminate or, at least, to lessen the amount of 'noise' introduced after the interviewer has parted with the interview schedule, we have been forced to perform a great deal of calculating to establish the data record from which deviations have been measured. And in this context, a lot of useful (and sometimes depressing) information has been gained. As already touched upon, this method of evaluation is time-consuming, and costly.

Questions vary considerably in degree of consistency. It is worth noting the results concerning the respondent's own socio-economic status and that of his/her father's during the respondent's adolescence.

Socio-economic status is coded from these answers on occupation. *g* and *M* refer to a dichotomization into blue collar workers and those otherwise occupied. For the complete table (8 status categories), the per cent deviating from previous responses was 9.5 and 22.1, respectively. The results reflect the different demands put on respondents in the two cases. This ought to be of interest for people studying social mobility.

There is another observation for father's socio-economic status, viz. from the regular SLC. A sample from 1976 was again used in the SLC in the spring of 1983. Thus, answers to a question which should give the same results in 1976 and 1983, when the same individuals are used, can be compared[4]. It was found that $g=15.6$, $M=.32$ and the per cent deviating was 29.7. A significant difference in marginals was obtained, meaning that in 1983 fewer interviewees placed their fathers in the blue collar status. Otherwise, the results concur, one point being that the 1976 SLC sample was reinterviewed in 1983 by the same method, i.e. personal interview.

The use of telephone interviews in the reinterview studies will be discussed later on. It may be noted that in his analysis of panel data from the U.S.A. and Britain, Schreiber (1975–76) found about the same inconsistency (per cent deviating) for 8 categories of father's occupation.

A summary of the 1979 reinterview study divided the survey questions into three groups:

i) The first group consisted of questions where the answers had a high consistency rate. This generally meant simple questions about concrete facts, e.g. questions about whether there was a kitchen, running water, drains, a shower, W.C., and the like in the home. Uncomplicated questions about educational background also belonged to this group, as well as questions on employment – whether the respondent had been working full-time or part-time, been on leave, been running a farm, etc. – and about the job itself – number of hours worked, travel time to work, income, etc.

ii) Questions classified in the second group had a medium rate of response consistency. The second group included some questions with somewhat vaguer definitions, on whether the work was safe and secure, whether it was noisy, whether it called for physically strenuous, or otherwise unsuitable working postures, etc. This category also included some questions about health which seem reasonably exact, such as whether the respondent could run 100 metres.

The generally phrased question on whether the respondent judged his state of health as good or bad also belonged to this group.

[4] The question reads as follows: "What was your father's principal occupation or employment during your early years, i.e. before you reached the age of 16?"

iii) The questions of the third group were those with a very low rate of response consistency. These questions were generally of a vague nature. Examples are the questions about discomfort at home because of the cold, air pollution, etc. In this group were also questions about whether the respondent found it difficult to get started in the morning, felt noticeably tired during the day or in the evening, etc. This is not surprising, since these last questions must be influenced by temporary changes. The third group also included some vague questions about work: whether the respondent found his work monotonous, whether he felt uncomfortable in his work because of heat, cold, draught, insufficient ventilation or illumination, gas, dust, smoke or smog. Other questions that belonged here were whether the respondent was worried about his own or his family's economic situation or about the risk of accidents at work.

One can compare with the conclusions drawn by Hochstim and Renne (1971) in their reinterview study. According to those authors, the reliability varied considerably according to the type of question, and was highest when objective facts were involved and lowest when feelings and attitudes came into play.

2.3.3. Some reflections on the importance of shifting data collection method and other confounding factors

There are a number of confounding factors to discuss before we can leave the problem of equivalence of regular interviews and reinterviews in our studies, that is changes in:

i) Method of data collection, mainly from personal to telephone interviews. In the 1979 study, the original interviews were carried out as personal interviews, while in the reinterviews 346 interviews were made by telephone and 158 by mail survey.

ii) Interview situations. The regular interviews were done by interviewers "in the field." Except in the 1984 study, all reinterviews have been performed by a centralized interviewer group. The two groups may differ in selection, training, and/or homogeneity of behaviour.

iii) Context. By sampling questions from the regular interview, contextual effects may occur.

iv) Surprise value. We use this term to signify that survey questions vary in the extent to which they surprise you as a respondent. They may surprise you because you have never thought about the issue before or your knowledge is stored in a different context in your memory from that supposed in the question. The second time you meet the question you are less surprised. Effects of (less) surprise are independent of memory effects. You may remember at the reinterview what your answer was in the original interview but still change it due to reconsideration because the question no longer surprises you. This effect may show itself both in systematic and random shifts.

v) Living conditions. The reinterviews have contained questions to uncover actual change between measurements, but they undoubtedly vary in ability to do this.

Evaluating these confounding factors means asking whether they exaggerate the inconsistencies in survey questions. We are prone to disregard point one and maintain that the particular contribution from the data collection method often has little effect on the result, as compared with the more decisive influence of many other factors, say whether the questions are concise or vague, or whether they are easy or difficult to answer. It may be noted that calculations of the average gross difference rate for the telephone interview and the mail survey part of the study, respectively, show identical percentages of 7.7. The results agree with

those reported by Pollard et al. (1978) in a study in which replicated measurements were done both by interviews and by self-administration.

As to interview situations, we can only say that it has possibly been at work. The fact that these centralized interviewers do much work in the follow-up part of the surveys may lead to interviewing behaviours differing from those in the field. Also, they may on the average resemble each other more. However, in the fourth study, from late 1984, interviewers in the field as well as the centralized group were used, and some information on the effect of this factor has been gained. Gross difference rates tend to follow each other in the two groups and the average levels are about the same.

Changes in context are inevitable since the reinterviews are much shorter. Effects of these changes are of course possible, but, we believe, subordinate to the influence of the fourth factor, changes in surprise value. Although this is mostly conjecture, we believe this factor to be important relative to the others. The main effect would then be assumed to lead to "better," more considerate answers in the reinterview. Few questions in the Survey of Living Conditions are of the type where spontaneous answering is desirable.

As to the fifth factor, we must acknowledge that questions about changes between measurements have weaknesses, like other survey questions. Still, we believe this factor does not seriously affect results.

Our overall conclusion is that these confounding factors are irritating when assessing question quality. Ordering questions according to inconsistency levels yields useful information. Exaggerations of inconsistencies due to these confounding factors may exist for some questions, such as that about prolonged illnesses, and must be borne in mind. For other questions, we may require that they should stand up to the test.

2.4. The Use of Reinterviews in the Labour Force Survey (LFS)

2.4.1. General outline of studies

The LFS is carried out monthly, by telephone interviews with a sample of 22 000. Each person in the sample is to be interviewed eight times over a two-year period.

In 1978 two reinterview studies were performed in order to get quantitative measures of the quality of the regular LFS (Bergman and Thorslund (1980)). One study comprised independent reinterviews with a sample of 632 individuals from the regular survey (non-response 6 %), the other reinterviews with reconciliation with a sample of 105 (non-response 2 %).

The emphasis was thus on reinterviews where the interviewers did not have access to answers from the regular interviews thereby avoiding the risk of interviewers being influenced by earlier answers. In reinterviews with reconciliation, this is a real risk which leads to underestimating errors. The risk must be weighed against the potentially valuable insights into error causes which reinterviews with reconciliation can give.

2.4.2. Some results

Labour Force status (LF-status) is a central concept in the LFS. The questionnaire is largely self-coding, and after only a few questions the interviewer can see from the questionnaire to which LF-status the interviewee belongs. As the interview proceeds, persons classified in different LF-statuses are asked different sets of questions.

Similarly, the LFS statistics are to a great extent presented according to LF-status. LF-status 4 – "Looking for work" – is a group of great interest to various users.

The number of persons who have been differently classified in the regular interview and in the reinterview is shown in Table 2.

Table 2. Reinterviewees by LF-status in the Regular Interview and the Reinterview

LF-status in the regular interview	LF-status in the reinterview 1	2	3	4	5	1–5	Deviations, in %
1 (outside the labour force)	99	2	–	2	10	113	12
2 (gainfully employed)	1	127	–	–	1	129	2
3 (absent from work)	6	15	96	–	–	117	18
4 (looking for work)	16	5	1	111	–	133	17
5 (incapable of work)	12	5	2	–	85	104	18

The table shows that there are substantial deviations between the classification in the regular interview and in the reinterview for all LF-statuses except LF-status 2 (gainfully employed during the week).

In order to estimate the frequency of deviating LF-status for the whole population, calculations were made taking into account that different LF-statuses were of varying frequency and that different persons had been sampled with different probabilities, depending among other things on region. As a rough estimate, it can be said that about 7 per cent of the Swedish population would be classified differently with respect to LF-status in a regular interview reinterview.

Assuming that the reinterviews had on the whole provided more valid answers[5], the following statements for the whole population could be made:

LF-status 1	was overestimated by	11 %
LF-status 2	was underestimated by	1 %
LF-status 3	was overestimated by	19 %
(LF-status 4	was underestimated by	3 %[6])
LF-status 5	was underestimated by	46 %

The results undoubtedly show that LF-statuses differ in ease of classification.

[5] The re-interviews ought to be more reliable than the regular interviews considering that:

I) As in the SLC reinterviews, all the reinterviews were made by a centralized group of experienced interviewers, who had been given special training for this task, including instructions for the treatment of borderline cases between various LF-status groups.

II) The interviewers making the reinterviews were given generous time limits, and the importance of meticulous interview work was strongly emphasized as well as the objective of the study. This should be compared with the working situation for the interviewers in the regular LFS work in the regular LFS work, where at this time the job was paid according to a piece work system (price per interview).

III) In addition to a meticulous LFS interview, the reinterview included some additional questions to make the LF-status classification as reliable as possible.

IV) In the reinterview, questions were included about the situation during the week before and the week after the measurement week. This was done both to eliminate the risk of misunderstandning which week was actually the measurement week, and to reduce the risk of the interviewee answering according to a normal situation instead of reporting the actual situation of the measurement week.

[6] The difference is not statistically significant, as the standard deviation is large for the national estimate of the percentage of reinterviewed persons belonging to LF-status 4.

2.4.3. On the use of reconciliation

The reinterviews with reconciliation were carried out in two steps:

i) The interviewer carried out an LFS interview, following the questionnaire meticulously. The interviewer had before him/her the interviewee's answer in the regular LFS, and if the interviewee gave an answer which did not agree with the one given in the regular LFS, the interviewer was to try as tactfully as possible to find out the reason for the deviation and then register the correct answer. The interviewer was instructed to take abundant notes in such cases.

ii) The same supplementary questionnaire that was used in the study with the independent reinterviews was used in the reconciliation interviews. In this part of the interview, the interviewer summarized in informal terms the interviewee's work situation during the measurement week and asked about the situation in the week immediately preceding and the one immediately succeeding the measurement week. This was done in order to minimize the risk of misunderstanding.

Reinterviews with reconciliation make heavy demands on the interviewer (in addition to the general quality requirements) if detailed information about the causes of deviations are to be obtained. Consequently, the re-interviews were mainly entrusted to a particularly experienced interviewer. As far as possible, the interviews were to be conducted with the sampled person himself/herself, and not by proxy.

The reinterviews with reconciliation, in which the interviewers could directly spot any discrepancies in the answers, provided results very similar to those of the independent reinterviews. Most often it could be established that the error had originated in the regular interview. For instance, the interviewee had answered "how things usually are" instead of how things actually were during the measurement week, or had had vague ideas about which week the questions pertained to, or the interviewer had made mistakes, etc.

In the reconciliation interviews, the interviewees themselves could decide which one of the two different answers was the correct one. The results in these cases seemed to confirm the conclusion about the greater validity of the re-interviews.

2.5. Some Comments on Efficiency

Quantitative measures derived from replications primarily as "alarms;" high values tell us to look harder for indicators of how the questions work. This alarm function may be fulfilled routinely in simplified, less expensive forms, if you lower your claims on realism and representativeness.

Returning to the studies discussed earlier in this section, we are forced to conclude that there is variable inconsistency in surveys like the Swedish SLC. This inconsistency is on the average considerable. This conclusion raises questions of possible quality improvements of data collection, such as giving up aspirations to measure every possible aspect of living conditions or improving the question form and definition or in other cases expanding the kind of data analysis possible. In the end, the survey analyst and user are not too interested in getting to know that data are "dirty;" they want instead to be able to make corrections and eliminate the effect of inconsistency. The manner in which the inconsistency measures can be transformed to 'error probabilities' will not be discussed here. But it seems clear to employ these techniques requires further evaluation of confounding factors in the reinterviews.

As to what can be done about eliminating these confounding factors, it must be borne in mind that the form of data collection in the reinterviews in the SLC, i.e., short telephone interviews by a centralized group of inter-

viewers, was actually the last resort. The alternative was no realistic reinterview study at all. For the amount of information they give, we find these reinterview studies inexpensive.

The reinterview method has been effective in revealing different weaknesses in the regular LFS. It became apparent that at many times the regular questions did not reflect the current situation. The prototype for the Swedish LFS questionnaire was drawn up in the U.S.A. in the 1950's. The questionnaire simply was not suited to today's more complicated situation (part time work, retraining, partial retirements, etc.). Thus , this led to changes being made in the interview form; the real gains were to be found in changing the form instead of the then current method of changing the instructions to the interviewers.

3. Systematically Scrutinizing Questions

3.1. Background

Survey questions are scrutinized more or less systematically, regardless of the methods used for 'field' testing. Handbooks on questionnaire design, from Payne (1951) to Sudman and Bradburn (1982), often contain a checklist through which a questionnaire may be desk-tested. In addition most questions are discussed in group sessions.

The work described in this part is not original except in its end result and ultimate aim.

With the 1979 Survey of Living Conditions, studies of change became of immediate interest. It was necessary to make a quality assessment of each variable that could be used for comparisons over time. It was in this connection that the above mentioned interview observations and the reinterview study of 1979 were performed.

It was difficult to give general recommendations for how the analysis should be carried out. Question quality varried so greatly that general rules were not applicable.

3.2. The Schedule for Question Scrutiny and Its Ultimate Aim: 'Measurement Profile'

A system for scrutinizing and assessing the quality of interview questions were developed. All questions from the 1979 SLC, which were also included in the 1975 survey, were scrutinized and assessed according to a schedule dealing with aspects of:

- variable definition
- question form and context
- respondent burden (e.g., amount of information retrieval and information handling required).

Some of the ideas for this schedule were obtained from Sudman and Bradburn (1974), that is aspects from the coding scheme used for classifying methodological studies to study response effects.

The end product of this scrutiny, which was done in group sessions, was a protocol on observed and perceived weaknesses for each question and an overall assessment of suitability for use in studies of change. Of all the questions, less than 40 % were dropped without notes of problems and in somewhat more than 10 % of the cases, the assessment unsuitable was used. In large part, the assessment was made that it is possible to study change but that caution should be used.

The ultimate aim of the schedule for classifying questions was to obtain measurement profiles for survey questions. Any survey question could be routinely scrutinized and classified according to the schedule and thus its characteristics described in numbers in a profile. Hopes for the use of such profiles will be discussed in Section 3.4.

But the ambition in using the scrutiny schedule had to be lowered. We could not fulfil the aim of establishing measurement profiles for all the questions, which would be quantitatively verified later by reinterview results.

3.3. Some Comments on Efficiency

Although we are concerned here with what is done more or less routinely in every survey, we feel that the extensive and systematic character of the question scrutinizing in the 1979 SLC merits some comment on efficiency. Clearly, the scrutiny had a general educational effect in sensitizing us to what to look for in questions. We did arrive at methods for evaluating a question's general quality and its potential usefulness in future change studies. Our method proved less quantitative than we had originally hoped.

Scrutinizing questions already in use and which have been tested in pilot surveys may appear problematic, especially if this is done in group sessions with the responsible question designer and/or analyst participating. However, that all the questions were scrutinized and fairly systematically made the task easier.

It was clear that further work was needed on the schedule, in particular in relation to classifying aspects of question issue.

3.4. Hopes for the Future: 'Profile' Use in Multivariate Study

The four reinterview studies in SLC comprise a considerable number of survey questions for which quantitative aspects of quality are available. Now, if these questions are classified according to a numerical measurement profile, the next step is naturally a multivariate study with questions as statistical units, the measurement profile as independent variables, and measures of the type used in Section 2 as dependent variables. One study of this type was made by Molenaar (1982), although his dependent variables were not directly derived from replications.

4. Validation Studies

4.1. Some General Remarks

Validation studies are often hard to perform. It is tedious to take one variable at a time and assess the extent to which it measures what it was intended to measure. There are often problems in finding an adequate criterion for validation studies.

We know from the replications presented earlier that survey questions on work environment vary considerably in reliability. Looking for the reasons for this, we touch upon validity aspects. For example, many of the results obtained in surveys where employees answer questions about their work environment reflect not only actual work environment but also their evaluations and demands. Results can be hard to interpret and less suited to absolute estimates of conditions for a particular group.

The reliability and validity of the results depend on whether you can reasonably use every day language for questions and answers in the field of work environment. Is it really meaningful to ask employees or are other measurement techniques necessary, such as technical measurements or ratings by experts, etc? And, of course, the problem of possible changes in the meaning of every day concepts when comparisons over time are concerned.

4.2. General Outline of Study of Work Environment

In everyday conversation, much is left implied because those taking part assume that they hold common definitions and experiences. Besides, there are benefits from establishing that implications are shared. In survey research, there are fewer benefits and one is easily tempted to leave much implied – for financial reasons alone. It may seem a truism to state that, other things being equal, the less left implied, the greater the validity and reliability of results. Nevertheless, we need to find out where to place the boundaries.

An ongoing methodological study at Statistics Sweden is attempting to assess the inherent vagueness of a number of interview ques-

tions pertaining to work environment[7]. Three types of questions have been studied:

i) Questions as to whether there is anything in the work environment that the employee considers a nuisance (e.g., whether the employee is 'bothered by noise').

ii) Questions as to how the employee would classify the work environment, the classification criteria being not altogether clear. For example, the question might be whether it 'is noisy'. Most questions on work environment in the SLC can be thus classified.

iii) Questions in which greater specifications are made (e.g., whether there is so much noise that it is impossible to carry on a conversation in a normal tone of voice).

Through the use of these different questions, the degree of precision in the communication could be controlled. The vagueness is greatest in the first type on questions. Neither 'noise' nor 'nuisance' is an unambiguous concept, and further difficulty is added by that the same noise may cause a wide range of reactions, depending on the sensitivity of the individuals concerned.

The second type of question also has inherent communication problems. The language does not provide an exact definition of the difference between a noisy and non-noisy environment, for example.

The third type of question is the least vague. Here one tries to make an operational definition of the matter of interest. We know ourselves what we are asking about, and the respondents understand the question they are to answer. Of course, even the third type of question may cause problems, as when a respondent is careless in his/her way of considering and answering the questions, or if he/he has memory problems. In addition to this, there are other reasons why a respondent may have trouble submitting a precise answer.

In the study of these survey questions, detailed descriptions were collected from number of places of work, and the descriptions compared with the answers submitted in mail surveys covering the same group of employees.

The detailed descriptions of the workplaces were obtained from the National Board of Industrial Safety, the Royal Institute of Technology in Stockholm, and others. The studies had covered such areas as noise, dust, and other substances in the air, temperature, draught, air circulation, vibrations and working positions, as well as medical data obtained from blood tests. Both technical measurements and other methods of obtaining descriptions had been used.

The variable 'noise' can be taken as an example of how the comparisons between a technical measurement and the result of a question were made. A positive answer to the question whether it is 'impossible to carry on a conversation in a normal tone of voice' was regarded as 'true' when the decibel level was 75 DB(A) or more. At a level of 100 DB(A) or more, the employee should agree with the statement 'of having to scream directly into the ear to be heard'.

The work environment concept covers much more than the items which can be studied by means of established technical methods. This means, for example, that is was not possible to study the validity of questions concerning a number of psychosocial aspects of the work conditions.

4.3. *Some Results*

In Table 3, some correlations between employees' answers and actual work environments, as decided by other evidence, are given.

[7] Further information can be obtained from A. Wikman, Statistics Sweden.

Table 3. Correlation between actual work environment and submitted answers about environmental problems, as measured by various types of questions (product-moment correlation)

	Discomfort questions (= type i)	Vague questions (= type ii)	More precisely formulated questions (= type iii)
Uncomfortable working positions		.76	.92
Heavy lifts		.67	.69
Repeated and monotonous operations		.44	.67
Exposed to cold	.51		.63
Exposed to draught	.55		.66
Exposed to vibrations		.68	.87
Exposed to noise	.48	.61	.73
Exposed to dust	.58		.66
Qualifications and skill required	.53	.66	.83
Mean correlation	.53	.64	.75

To conclude, the study has shown the possibility of increasing the validity and reliability of employee surveys by sharpening the question tools.

Questions which call for value judgments or are vaguely phrased tend to cause difficult measurement and interpretation problems. On the other hand, concrete questions (of type iii) give quite high correlations with more traditional measures.

The fact that the work environments included in the study had been subjected to thorough examinations makes it likely that the respondents had better knowledge and a higher degree of awareness with regard to their work environment than the average employee. The difference in validity between different types of questions may be even greater for an environment which had not previously been studied than for the environments included in the project.

4.4. Implications and Effects of the Results

Concrete questions, where the items of interest are clearly defined, generally present fewer problems. This concretization, however, often calls for a subdivision of the main question and a consequent increase in the total number of questions. If we ask about 'heavy lifts', one question may be sufficient. On the other hand, if we try to concretize, we may have to specify the weights and perhaps ask several questions about different categories of weights. It may also be necessary to ask how the lifting is done in order to find out how physically strenuous it is. The questionnaires then tend to become longer.

In surveys with a wide coverage like the Swedish SLC, which concern a great number of aspects, there is very little room for further questions in any particular field. It might therefore be difficult to include within the framework of these surveys all the specifica-

tions and follow-up questions needed to ensure greater quality. In other words, improved quality in one area might make it necessary to exclude questions in another.

Since the aim of the Swedish Survey of Living Conditions is to produce objective descriptions of reality, the findings of the study have led to changes. Some questions have been changed despite the possibility of comparison with earlier results may have been lost. But the argument for keeping comparisons over time weighs less in light of experiences especially with questions of type i) and, in part, with type ii).

The study has shown that it seems possible to get valid measures of some aspects of work environment by asking employees. But that questions of type iii) are limited in scope means that a lot of questions are needed to cover the whole picture. The aim of making a broad description of the work environment of the Swedish labour force possible can not be fulfilled within the SLC.

So, special surveys are needed. Among the choices is the possibility of using the LFS as a frame for sampling employees by occupation and following up by questions on work environment, partly by telephone, partly by mail survey.

4.5. Some Comments on Efficiency

The validity of questions on work environment in the regular Survey of Living Conditions has been studied and compared to the validity of more concretely phrased alternative questions. The results of these comparisons have led to some changes in the regular SLC. The SLC, like any continuing survey, is antagonistic to changes that jeopardize comparisons over time. In this case, the evidence of the results was clear; consequently a reduced potential for analysis was accepted.

But convincing results do not automatically mean that these kinds of studies are efficient. The study described above was favoured by the fact of some already existing data which could be used in validation, like technical measurements. So it was partly possible to lean on the work of others. If the cost and effort of all the data used in this study is included, the question of efficiency may have to be reconsidered. How many of the deficiencies could be found by simpler and less costly methods? The effect of varying the amount of vagueness in questions should show up even in consistency studies. But one of the advantages of validity studies is that they are easier to analyze. Variations in correlations between the questions studied and the 'truth' give clearer indications to act upon than you get in reinterview studies.

5. Discussion

Experiences with different methods of studying question quality in surveys have been presented, viz., observations of interviews, reinterview studies, systematic question scrutiny and validation studies. These methods have provided valuable insights and results. The methods differ as to what type of information they yield, and how difficult or costly they are. When resources are limited, as they mostly are, how do you make wise choices between different methods? Which is most efficient in detecting deficiencies?

One important aspect of our work, confirmed by schedule scrutiny and the reinterview studies, is the fact that the largely 'atomistic' outlook of survey analysis conceals many errors. Whereas when patterns of answers are studied, new insights are gained. By this we do not mean that all inconsistencies found are really errors – only that you see better.

Efficiency of different methods, in a general sense, is concerned with the number and kinds of deficiencies discovered. This can be seen both in a relative and an absolute sense. In an

absolute sense, you need to turn to experiments to answer questions like 'How many of a planted set of deficiencies are discovered by this method'? Our views have been presented in a relative sense. This pertains to the view of conventional pilot methods as less efficient than the methods discussed here. And to the kinds of deficiencies the discussed methods discover.

As to the aspect of level of abstraction in survey questions, we may note that the problem of finding a reasonable level cannot be satisfactorily dealt with by desk methods – this must be done by field testing and evaluations. When methods of field testing are concerned, we think that a lot of reasons speak for the use of more than one method. That is, that measurement and understanding are needed to balance each other.

Both in measurement and understanding approaches, the competing demands within survey research are revealed. Choosing a high level of abstraction in survey questions may reduce quality, choosing a low level may mean losing the chance to cover many interesting subjects. The interviewers employed by Statistics Sweden have participated in primary training and later training programmes, have extensive experience on the average and, even seen in an international perspective, must be categorized as well suited for their job. Still, the results from observations and from reinterviews, especially those with reconciliation, show what we are now prone to call self-evident – that interviewers have shown many different behaviours. Interviewers need training, but the basic survey problems cannot be trained away. Our conclusion is that the inherent weaknesses in survey research must be more widely acknowledged. Measuring instruments used should be less sensitive to human weaknesses.

Validation studies can give very clear evidence on how different question principles work. And it could seem almost self-evident that validation studies are conducted when planning a new survey. Yet, in the sense of studies where external 'evidence' is used to scrutinize survey questions, validation studies are rather rare, both in pilot studies and as controls of continuing surveys. They are complex and costly and it is sometimes difficult to specify the external evidence to use for validation.

The evidence you get, once validation is completed, is often very valuable. But from an efficiency point of view, a validation study is perhaps not always the best choice. It seems that part of the deficiencies observed in validation studies can be traced by simpler and less expensive methods. And since lack of reliability will lessen validity, it follows that sometimes priorities between testing/assessing methods may be established.

To conclude, reinterview studies have been found to efficiently rank survey questions according to quality. You can get quantitative measures (independent reinterviews) as well as understanding (reinterviews with reconciliation) relatively smoothly.

In question design and analysis, you work from your sense of how everyday language functions and from experiences of how similar questions worked previously. This process is automatic and intuitive. It seems to us much is won by supplementing intuition with a systematic perspective. Systematic scrutiny of questions works as an antidote to going off hastily. Again, it would seem to us to give hints for priority discussions.

Furthermore, if coding questions into a measurement profile is conducted routinely and if those quantitative aspects of question quality which may be available are added, a data bank for methodological research will sooner or later be available. The schedule for question coding which was developed for and used in the Survey of Living Conditions needs further improvement, however, before multi-

variate studies of factors affecting question quality can be performed.

This is one of the ways to increase our knowledge of how different factors affect question quality, that is further studies that make it possible to order them according to importance.

6. References

Belson, W. A. (1963): Studies in Readership. Business Publications Ltd., London.

Belson, W.A. (1968): Respondent Understanding of Survey Questions. The Reprint Series of The Survey Research Centre, No. 40. The London School of Economics and Political Science.

Belson, W. A. (1981): The Design and Understanding of Survey Questions. Gower Publishing Co. Ltd.

Bergman. L. R. and Thorslund, M. (1980): Response Quality in the Swedish Labour Force Surveys – Findings of Two Reinterview Studies. Methodological Studies from the Research Institute for Statistics on Living Conditions, No. 13 E, Statistics Sweden.

Bohrnstedt, G. W. (1983): Measurement. In Rossi, P. H., Wright, J. D. and Andersson, A. B.: Handbook of Survey Research, Chapter 3. Academic Press, New York.

Cannel, C. F., Marquis, K. H., and Laurent, A. (1977): A Summary of Studies of Interviewing Methodology. Vital and Health Statistics, Series 2, No. 69. U.S. National Center for Health Statistics.

Christoffersen, M. (1984): The Quality of Data Collected at Telephone Interviews – Investigations of Differences in the Quality of Surveys Conducted by Personal and Telephone Interviewing. Statistisk tidskrift (Statistical Review) nr 1, pp. 27–35.

DeMaio, T. J. (ed.) (1983): Approaches to Developing Questionnaires. Statistical Policy Working Paper 10, Statistical Policy Office, Office of Information and Regulatory Affairs, U.S. Office of Management and Budget.

The General Household Survey (1973): Introductory Report. Office for Population Censuses and Surveys, London, HMSO.

Hochstim, J. R., and Renne, K. S. (1971): Reliability of Response in a Sociomedical Population Study. Public Opinion Quarterly, Vol. XXXV, pp. 69–79.

Hunt, S. D., Sparkman, R. D., and Wilcox, J. B. (1982): The Pretest in Survey Research: Issues and Preliminary Findings. Journal of Marketing Research, Vol. XIX, pp. 269–273.

Molenaar, E. M. (1982): Non-Experimental Research on the Effects of the Wording of Questions in Survey Interviews. Quality and Quantity, 16, pp. 69–90.

Payne, S. L. (1951): The Art of Asking Questions. Princeton University Press, Princeton, New Jersey.

Pollard, W. E., Bobbitt, R. A., and Bergner, M. (1978): Examination of Variable Errors of Measurement in a Survey-Based Social Indicator. Social Indicators Research, 5, pp. 279–301.

Rodgers, J. L., Billy, J. O. G., and Udry, J. R. (1982): The Rescission of Behaviors: Inconsistent Responses in Adolescent Sexuality Data. Social Science Research, 11, pp. 280–296.

Schreiber, E. M. (1975–76): Dirty Data in Britain and the U.S.A.: The Reliability of 'Invariant' Characteristics Reported in Surveys. Public Opinion Quarterly, Vol. XXXIX, pp. 493–506.

Sletto, R. F. (1940): Pretesting of Questionnaires. American Sociological Review, Vol. 5, pp. 193–200.

Sudman, S. and Bradburn, N. M. (1974): Response Effects in Surveys. Aldine Publishing Company, Chicago.

Sudman, S. and Bradburn, N. M. (1982): Asking Questions. Jossey-Bass Publishers, San Francisco.

Thorslund, M. and Wärneryd, B. (1985): Methodological Research in the Swedish Surveys of Living Conditions. Problems of Measurement and Data Collection. Social Indicators Research, Vol. 1, pp. 77–95.

Vogel, J. (1981): Social Report on Inequality in Sweden. Living Conditions Series, Report No. 27, Official Statistics of Sweden, Stockholm.

Vogel, J. (1982): The Swedish Annual Level of Living Surveys: Social Indicators and Social Reporting as an Official Statistics Programme. Paper presented at the 10th World Congress of Sociology, Mexico City, 16–20 August 1982 (Obtainable from Statistics Sweden).

Received August 1984
Revised April 1985