7

Why Do Surveys Disagree? Some Preliminary Hypotheses and Some Disagreeable Examples

Charles F. Turner

Overview

This chapter reports a number of anomalies in the survey measurement of subjective phenomena. Initial examples are drawn from recent attempts to use survey measures as social indicators of phenomena such as the well-being of the population, public confidence in national institutions, and public support for science. Observed discrepancies between supposedly equivalent measures of the same phenomena indicate that survey measures of many subjective phenomena have large nonsampling variances that may contaminate both their univariate and multivariate response distributions. These nonsampling variances may confound attempts to measure population change in instances where complete survey replication is not performed. This is true because variation due to change in the population

cannot be disentangled from variation due to changes in the measurement process itself.

Our present lack of systematic knowledge about the effects of factors, other than sampling, on the survey measurement of subjective phenomena often makes it difficult to predict which aspects of a survey must be replicated (and which can be ignored) to produce comparable measurements. This chapter reports results of several experiments involving systematic variations of survey contexts. While these experiments demonstrate the presence of some large effects due to experimental variations in survey context (for example, 14 percent fluctuations in the univariate response distributions), the unanticipated nature of many of these findings and the inconsistency of the results themselves testify dramatically to our inability to anticipate the impact of common variations in survey procedure.

While such problems are not rare, we nonetheless simultaneously observe many instances of apparently robust measurements of subjective phenomena. We suggest here that there may be organizing principles concerning the types of subjective measurements that are particularly vulnerable to nonsampling artifacts. A review of recent instances of replicated survey measurements provides modest but incomplete support for this approach.

The need for research on the components of measurement variability attributable to sources other than sampling is discussed. It is suggested that this research should involve coordinated methodological studies conducted across a number of survey organizations and should prompt a fundamental reconsideration of the psychological assumptions that underlie the practice of survey research.

AREA OF INQUIRY

Traditionally the term "subjective" has been used to denote those phenomena that are, in principle, directly observable only by subjects themselves. Phenomena of this sort include those commonly labeled attitudes, beliefs, and opinions. These may be conceptually distinguished from other phenomena which, although frequently measured by subjective means (that is, self-report), are theoretically amenable to objective (that is, independent) confirmation.[1]

For example, while we may measure age by asking respondents to report it, it would be theoretically possible to obtain independent evidence from other witnesses. For this reason we would not label chronological age, per se, as a subjective phenomenon. In theory, many other phenomena may also be measured independently of a subject's own report (for example, educational attainment, geographic mobility, fertility history, family structure, income, and so on). However, many important phenomena are inherently subjective and thus immune to independent verification. In particular, we have no direct knowledge of an individual's "attitudes," "beliefs," or "opinions."[2]

The present inquiry focuses upon survey measurements of such subjective phenomena.

USE OF SUBJECTIVE SOCIAL INDICATORS

Traditionally, national statistics have been the domain of demographers and economists. Inquiries made by the U.S. Bureau of the Census have generally been limited to assessments of the size and distribution of the population, and a variety of other phenomena that are, at least theoretically, amenable to independent corroboration, for example, age, income, educational attainment, and so forth.[3] This does not mean, of course, that subjectivity does not contaminate such assessments; the use of self-report inevitably raises this issue. However, the validity and reliability of survey estimates of such phenomena may be evaluated by using independent data (for instance, birth and earnings records) to estimate the magnitude and sources of error (and bias) introduced by the exclusive use of information supplied by subjects themselves. In contrast, the very concept of "true value" employed in such studies is difficult to conceptualize when one is discussing the measurement of subjective phenomena (Waksberg, 1975).

Our concern in the following pages centers upon examples of measurements that have been used in recent years as subjective social indicators. It would, however, be a gross oversimplification to distinguish between the "old objective" and the "new subjective" indicators. Some well-known measurements, such as the national unemployment index, contain fundamentally subjective elements.[4] Similarly, recent evaluations of the National Crime Surveys (Cowan, Murphy and Weiner, 1978; Gibson et al., 1978) point to the subjective components of crime victimization statistics.[5]

Nevertheless, in recent years national statistics have come to include an important and rapidly growing complement of statistics designed to measure explicitly subjective phenomena. For example, the Social Indicators program (Executive Office of the President 1973; U.S. Department of Commerce 1977; Bureau of the Census 1980) incorporates measurements of a wide range of subjective phenomena.[6] A recent volume of the *Social Indicators* series argues that such measures provide a

> . . . vitally needed supplement to traditional national statistics. The basic reason for including such subjective measures in this report despite the difficulties in their interpretation is that they offer a vital dimension in developing a comprehensive description of the condition of our society and the well being of its mem-

> bers. The bulk of the information presented [in this report] relates to people's objective situation or condition—their jobs, their incomes, their health status, etc. The main purpose of the attitudinal measures is to provide some insight as to how people perceive certain aspects of these conditions. Such data are an essential source of information. . . . [U.S. Department of Commerce, 1977:XXVI]

For similar reasons, the National Science Board's recent series of reports (1973, 1975, 1977) on the state of science in the United States has incorporated a concluding chapter on public attitudes toward science and technology. Interest in this topic follows from the assumption that financial support, the imposition of legal constraints (for instance, regulation of recombinant DNA research), and the recruitment of young people into the scientific professions depend in part upon public perceptions of science.[7]

The increasing importance of measures of subjective phenomena in federal statistical programs is paralleled by a growing range of relevant research activities in the academic community. This work has included psychological studies of well-being (for example, Campbell, Converse, and Rodgers, 1976; Andrews and Withey, 1976; Bradburn, 1969; Staines and Quinn, 1979), investigations by sociologists of trends across time in sex role stereotyping and the tolerance of nonconformity (for example, Davis, 1975a; Duncan, 1979; Mason, Czaijka, and Arber, 1976), and work by economists on the relationship of economic development to individual happiness (for example, Easterlin, 1974). In addition to such substantive work, considerable resources have begun to be invested in providing regularly replicated survey measurements to facilitate the study of social change (for example, the National Opinion Research Center's General Social Survey begun in 1972).[8] The wide dissemination of these rich data in the research community foreshadows the increasing use of such survey measures for scholarly research in the social sciences.[9]

COMPARABILITY OF SUBJECTIVE SOCIAL INDICATORS

Survey measurements of subjective phenomena are made by many organizations. In the United States, sources outside of the federal statistical system produced the majority of the subjective social indicator measurements reported in recent federal *Social Indicators* publications.

Use of data from a variety of sources inevitably raises questions of comparability. Despite one's hopes, comparability of measurement does not occur naturally. For example, in the natural sciences there has been a long history of concern with the difficult problem of ensuring the replicability of chemical and physical measurements. Early examples include the discovery of systematic variations in the observations of individual astronomers (see Boring, 1950). More recent work has included attempts to partition the variability in laboratory measurements in analytical chemistry into components representing the effects of different analytic procedures, measurable aspects of laboratory environments, interexperimenter differences, and differences between unmeasured attributes of laboratories that produce constant biases.

The experience in the natural sciences suggests that comparability of measurement is the result of careful standardization of research procedures, frequent calibration, and the continuous monitoring of performance. Equally clear is the fact that comparability is not easily achieved (Boffey, 1975), and thus published measurements of elemental physical constants sometimes reveal frightening discrepancies (see, for example, Hunter's [1977] plot of reported values for the thermal conductivity of copper, and data for other metals reported by Ho, Powell, and Liley [1974]). Such problems have led to the standardization of research procedures and the development of methods for collaborative tests among laboratories in such fields as analytical chemistry (Youden, 1975; Steiner, 1975). We will subsequently argue that problems of comparability in the survey measurement of subjective phenomena should prompt a consideration of parallel techniques by those interested in the development of more reliable subjective social indicators.

The Problem. The increased use of replicated time-series of subjective social indicators has spawned some disagreeable progeny. Most irritating has been the multiplication of instances in which supposedly comparable measurements have differed both substantially and significantly between surveys (see Turner and Krauss, 1978; Martin, 1983). Discrepancies of 15 percentage points have been observed between the univariate distributions of purportedly equivalent measurements of some subjective phenomena. Such large discrepancies prompt a number of questions. One would like to know, for example:

- What causes these measurements to disagree?
- Are these disagreements symptomatic of a larger problem or are they restricted to a few isolated cases?
- Are there any organizing principles which could identify indicators that are more (and less) likely to produce discrepant results?

In the following pages we review several examples drawn from current social indicators projects. We use these examples to illustrate the issues and to demonstrate the need for further research on the nonsampling components of variance in the survey measurement of subjective phenomena. We

also propose some preliminary ideas concerning the types of subjective indicators that are particularly vulnerable to artifacts of measurement. We do so not in the hope of formulating final principles, but rather to provide initial hypotheses around which future research might be organized.

As a starting point, Turner and Krauss (1978:468) suggested that the vulnerability of measurement to nonsampling artifacts might be a function of the survey questions themselves and of the phenomena they intend to measure. In particular, it was hypothesized that vulnerability would be concentrated among survey questions that[10]

1. Were most amorphous in their meaning, for example, those seeking to assess "confidence," "trust," and so forth;
2. Were most ambiguous in their referents, for example, those inquiring about the "people running organized religion," and so forth;
3. Involved the most arbitrariness in the selection of a response category, for example, great deal versus some confidence; and
4. Dealt with topics that do not have a well-defined place in public discussions, for example, public evaluations of science.

While we argue in the conclusion to this chapter that these organizing principles are a reasonable place to begin, we do realize that they are at present overly general and are open to a wide range of interpretations. After considering some examples of problematic measurements, we attempt to show some ways in which these principles might be made sufficiently concrete to be tested in future research.

It should also be recognized that, when applied to most common measurements, the foregoing principles are often redundant. For example, questions about amorphous phenomena frequently and perhaps inevitably require arbitrary choices between response categories. At present we are unable to guess whether each of the foregoing characteristics of survey questions is equally important in determining a measurement's potential vulnerability to nonsampling artifacts. Indeed it is quite possible that it is the coincidence of two (or more) of these characteristics that induces vulnerability.

Readers will find no definitive answers in the following pages. The presently available evidence is too sparse and problematic to permit any confident disentanglements of the various effects. Methodological experiments will have to be embedded in future surveys to answer these questions. In the interim, we intend to illustrate the problems that are induced by the nonsampling components of the variance in survey measurements, to provide some insights into the sources of this variability, to document our considerable ignorance, to suggest the pressing need for research, to propose some avenues of inquiry, and to provide some examples of how one might proceed.

Examples

EXAMPLE I: DISAGREEMENTS ABOUT HAPPINESS

In the fall of 1977, we began an investigation of responses to national survey questions on personal "happiness." These questions have been incorporated in research attempting to define the nature of social well-being and to produce social indicators of life-satisfaction (for example, Gurin, Veroff, and Feld, 1960; Bradburn, 1969; Campbell, Converse and Rodgers, 1976). All of this work has gone beyond the notion that responses to a single question are ideal measures of subjective well-being. Nonetheless, responses to the simple question,

> Taken all together how would you say things are these days—would you say that you are very happy, pretty happy, or not too happy?

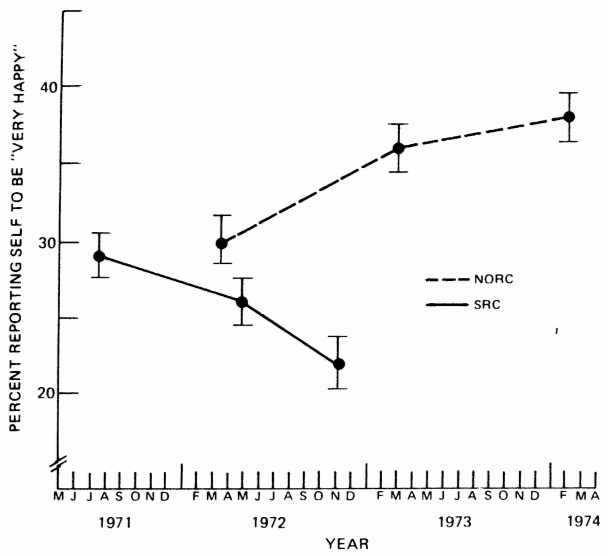
have been tracked from the year 1957.[11] And both the trends across time and differences between nations in response to this question have been analyzed by several authors (for example, Easterlin, 1974; Davis, 1975b; Campbell, Converse, and Rodgers, 1976; Andrews and Withey, 1976). Moreover, the responses to this "happiness question" have been used as a validity criterion in the development of more elaborate indices of life-satisfaction. Thus, responses to this question are of substantial importance and interest in their own right.

Figure 7.1 presents two independent series of happiness estimates derived from surveys conducted by the Survey Research Center (SRC) of the University of Michigan and the National Opinion Research Center (NORC) of the University of Chicago. It will be seen from this figure that there are not only discrepancies in estimates of absolute levels of happiness, but also that the (apparent) trends in the two series diverge. One series shows an apparent increase while the other series registers a decline in happiness.

We first noticed this "disagreeable" result in the fall of 1977, and it was the subject of preliminary discussions with an ad hoc working group which met to discuss discrepancies observed in the "confidence in institutions" series.[12] Subsequent examination of the two happiness series has caused us to doubt the validity of the comparison shown in Figure 7.1. Examination of the questionnaires used by NORC and SRC reveals a slight difference in question wording: the SRC version repeats "these days" after listing the response categories, while the NORC version does not (see caption to Figure 7.1 for exact wordings). Thus, it might be argued that the two questions were indicators of slightly different phenomena, although admittedly we might expect the trends across time to be parallel rather than divergent. On the other hand, it could be that the divergences are more apparent than

FIGURE 7.1

Trends in Self-Reported Happiness, 1971-73.



SOURCES: NORC, National Data Program for the Social Sciences: Codebook, 1972-74. SRC estimates from Campbell, Converse, and Rodgers (1976); survey dates from Campbell, Converse, and Rodgers (1976), Andrews and Withey (1976), and J. Varva (SRC, personal communication).
NOTE: Estimates are derived from sample surveys of noninstitutionalized population of the continental United States aged 18 and over. Error bars demark ±1 standard error around sample estimate.
QUESTIONS: "Taken all together, how would you say things are these days—would you say that you are very happy, pretty happy, or not too happy?" [NORC];
"Taking all things together, how would you say things are these days—would you say you're very happy, pretty happy, or not too happy these days?" [SRC]

real. For example, national happiness might have fluctuated rapidly between 1971 and 1973, and thus the data could be reliably monitoring month-to-month changes that were taking place in the national population.

While such arguments can be made, they are apologies rather than explanations. The NORC and SRC data have been treated as a unitary time-series by several authors, despite differences in wording (see, for example, Campbell, Converse, and Rodgers, 1976:26; Andrews and Withey, 1976: 319). Moreover, large month-to-month fluctuations in these indicators would preclude use of the (approximately) annual and biennial reporting schedules that have been the rule in the social indicators field.[13]

Despite their limitations, the data shown in Figure 7.1 prompted us to speculate about causes of the observed discrepancies. Our initial hypothesis followed from the fact that NORC altered its questionnaire in 1973 so that a question about marital happiness,

> Taking things all together, how would you describe your marriage? Would you say that your marriage is very happy, pretty happy, or not too happy?

immediately preceded the general-happiness question. In 1972 the general-happiness question followed questions asking about respondents' satisfaction with their financial situation and whether their financial situation had been "getting better, getting worse, or . . . stayed the same" in the last few years.

We hypothesized that insertion of this marital-happiness question created an artifactual response effect. Our initial examination of this hypothesis (see Table 7.1) indicated that:

- There was a high correlation (gamma = +0.75) between responses to the marital- and general-happiness questions;
- The marital-happiness question elicited a relatively high proportion (0.6) of "very happy" responses; and
- The increase in overall happiness between 1972 and 1973-74 in the NORC series occurred only among married persons (see Table 7.1).

This last finding[14] was particularly important because the hypothesized context effect could only have occurred for married individuals (unmarried persons could not, of course, be asked about the happiness of their marriages). On subsequent examination, we have also found that there was a

TABLE 7.1

Variation in Percentage of Married and Unmarried Respondents Reporting Themselves to Be "Very Happy" Between 1972 and 1974

	Year			X^2 for Temporal Change[a]	
Sample	1972[b]	1973	1974	1972 vs. 1973 vs. 1974[c]	1972 vs. 1973 + 1974[d]
Married	33.5%	42.7%	44.6%	22.0, $p < .001$	21.5, $p < .001$
Not married	17.9	20.0	19.6	0.4, *ns*	0.4, *ns*
Total	29.7	36.8	38.4	20.1, $p < .001$	19.6, $p < .001$

[a]Chi-square statistics were adjusted for design effects of NORC's clustered sample design by using a deflated sample size ($N' = 0.66N$) in computations. (Analysis of the intracluster correlations [median 1973-78 $r_i = 0.02$] for the happiness item indicates that this correction is not an unreasonable allowance.) This deflation in sample size makes it less likely for our tests to find significant differences between the results obtained by different surveys.
[b]1972 NORC General Social Survey did not include question on marital happiness.
[c]d.f. = 2.
[d]d.f. = 1.

significant alteration after 1972 in the association between responses to the general-happiness and financial-status questions. In later years, from 10 to 44 questions were interspersed between these items.[15]

While any comparison of the NORC and SRC happiness series admits to a plethora of alternative explanations (for example, wording effects, "house" effects, short-term temporal variations, and so on), the results of our initial explorations encouraged us to seek a better test for our hypothesis. We were fortunate to discover that a wealth of information on happiness was collected during this period. Between April 1973 and May 1974, NORC, with the support of the National Science Foundation (RANN division), conducted a series of pilot surveys to provide continuous monitoring of public opinion for policy makers in eight federal agencies. At intervals of approximately one month, NORC drew samples of the national population for interview. While the content of the surveys varied from month to month, the happiness item was included in every cycle of NORC's Continuous National Survey.

These data allow us to compare responses across time for two identically worded questions in surveys conducted by the same research organization. This comparison provides a control for both wording differences and any possible organizational idiosyncracies (for instance, variations in interviewer training). We obtained a copy of these data during February 1978 and set to work examining the plausibility of our context hypothesis for the misbehavior of the happiness time-series.

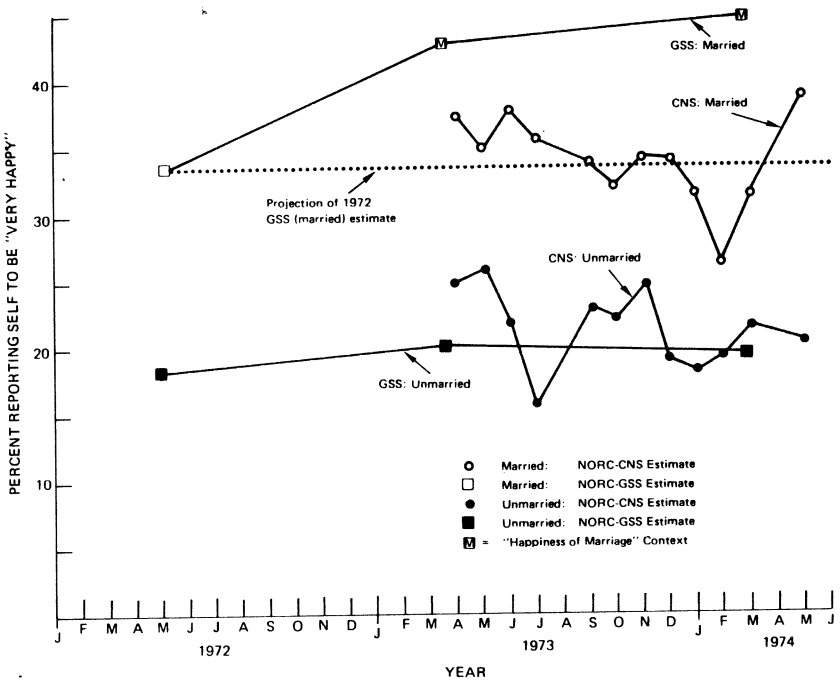
A graphic summary of our findings is presented in Figure 7.2. Specifically, we found that for unmarried individuals, yearly estimates derived from the NORC General Social Survey (GSS) and the monthly estimates from the NORC Continuous National Survey (CNS) were in general agreement. This is not to say that the estimates were identical. However, observed discrepancies were within the range expected on the basis of sampling error. In short, unmarried men and women responded to the happiness question in the same manner in the 1972, 1973, and 1974 General Social Surveys and the 12 cycles of the Continuous National Survey.

For the married respondents, a rather different result emerged. In particular, while the GSS happiness estimates exhibit a sharp rise between 1972 and 1973–74 (change = + 10 percentage points; Pearson $\chi^2 = 21.5$, d.f. = 1, $p < .0001$), the monthly estimates derived from the Continuous National Survey evidence no similar trend. Moreover, the CNS happiness estimates are consistently below those of the GSS. Indeed, as Figure 7.2 shows, the GSS measurement in 1972 (when the marriage question was not included) provided a better prediction of the CNS estimates in 1973 and 1974 than did the actual GSS estimates in those years.

Although other hypotheses might be supported, we concluded that

FIGURE 7.2

Variations in Response to NORC "Happiness" Question for Married and Unmarried Respondents in the General Social Surveys (GSS) and Continuous National Surveys (CNS).



NOTE: Estimates are derived from samples of approximately 1,000 (GSS) and 440 (CNS) married respondents and 500 (GSS) and 220 (CNS) unmarried respondents.

1. The internal evidence of a temporal trend only for married GSS respondents; and,
2. The predictability of the 1973–74 CNS data from 1972 GSS data provide strong support for the hypothesis that a response effect arose from the insertion of the marital-happiness question into the 1973 and 1974 General Social Surveys.

Experimental Evidence. It was subsequently possible to test our conclusions experimentally. Four experiments using the happiness questions were conducted by the *Washington Post* poll, the General Social Survey of the National Opinion Research Center, the Survey Research Center, and the Opinion Research Corporation of Princeton, New Jersey. Each of these

experiments manipulated the context in which the general-happiness question was asked.

The first three experiments (conducted by NORC, SRC, and the *Washington Post* poll) were similar in design. In one experimental condition in each of these experiments, the immediate context in which the general-happiness item was asked was "controlled" by asking it immediately after the question on marital happiness. In the other experimental condition, the context in which the general-happiness question was asked was uncontrolled; that is, it occurred after whatever other questions (varying from experiment to experiment) were contained in the questionnaire. So, for example, in the *Washington Post* survey, the general-happiness question followed a question on income in the uncontrolled context; in the SRC experiment, it followed a series of items on the gas shortage; and in the NORC experiment, it followed a series of five questions in which respondents were asked "how much satisfaction" they received from their cities or places of residence, hobbies, family life, friendships, health and physical condition.[16]

A parallel variation occurred serendipitously for the marital-happiness item in these experiments. In one experimental condition, it too followed these varying items, and in the other condition, it followed a fixed item: the general-happiness question.

As before, all samples were restricted to married respondents. Samples were drawn from the adult (aged 18 and over) population of the continental United States. The SRC and *Washington Post* experiments were done in telephone surveys using random-digit dialing to sample phone numbers. The NORC surveys were conducted as face-to-face interviews using a multistage area probability sample (see National Opinion Research Center [1980] for description of sample design).[17]

Figure 7.3 shows the proportion of respondents who said they were "very happy" in response to the marital- and general-happiness items for the controlled and uncontrolled contexts in each of the three experiments. Two important conclusions are evident to the eye in these plots:

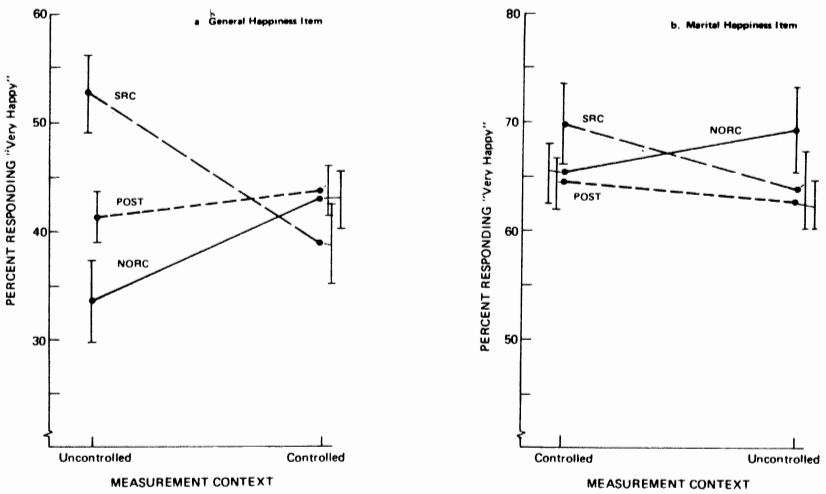
1. When the survey context was controlled, all organizations obtained similar estimates of the proportion that was "very happy" for both the general-happiness and marital-happiness measurements.

2. However, when the measurement context was left uncontrolled, the general-happiness measurement showed considerable volatility, while the more specific question on marital happiness appeared to be unaffected.

More formally, we note that we can fit a model (see Table 7.2) to the data from the marital-happiness measurements which fits only the univariate response distribution for the marital-happiness measurement; the observed

FIGURE 7.3

Percentage of Respondents Saying that They Were "Very Happy" in Response to Questions on General Happiness and Marital Happiness.



SOURCES: Data are from surveys conducted by the Survey Research Center, University of Michigan (SRC), the National Opinion Research Center (NORC), and by the George Fine Organization for the *Washington Post*. All samples were restricted to married respondents who had a telephone in their residence. See text for description of measurement contexts.
NOTE: Error bars demark ± 1 standard error for estimates.

variations across different measurement conditions and between different survey organizations are all within the range of sampling error (Likelihood-ratio chi-square: $L^2 = 10.8$, d.f. $= 10$, $p = .37$). In contrast, the general-happiness measurements show considerable variability in the uncontrolled context (the proportion saying "very happy" varies from 0.335 to 0.527), and require us to fit a three-way interaction term (Survey House by Context by Response) in order to obtain an adequate fit to the data obtained in this experiment.

We note that these results fit quite well the pattern originally hypothesized (the *general* question produced more labile measurements than the more *specific* item). It is also important to note that it would have been hard to predict in advance (and it is hard to understand after the fact) the direction and magnitude of the variations shown by the general-happiness measurement in the uncontrolled context. Thus, it is hard to intuit why SRC's measurements (which were made after respondents had been asked questions about the 1978–79 gas shortages) should produce an estimate of the proportion "very happy" that was so much higher than that of the *Washington Post* poll or the NORC experiment.[18] We do nonetheless note

TABLE 7.2
Test of Alternative Models for Behavior of Happiness Measurements

Model	Marginals Fit	d.f.	L^2	p
General-Happiness Measurements				
1. Stable measurements	{H} {CS}	10	28.0	.002
2. Context effect	{HC} {CS}	8	27.7	.001
3. Survey effect	{HS} {CS}	6	20.8	.002
4. Context and survey effects	{HS} {HC} {CS}	4	20.0	.001
5. Interaction effect	{HSC}	0	0.0	(n.a.)
Marital-Happiness Measurements				
1. Stable measurements	{H} {CS}	10	10.8	.37
2. Context effect	{HC} {CS}	8	10.7	.21
3. Survey effect	{HS} {CS}	6	5.9	.44
4. Context and survey effects	{HS} {HC} {CS}	4	5.5	.24
5. Interaction effect	{HSC}	0	0.0	(n.a.)

NOTE: Models were fit using procedures developed by Goodman (1971). L^2 values are likelihood-ratio chi-square statistics. Variables included in this analysis are
H = response to happiness question (three categories: very happy; pretty happy; not too happy). Respondents who did not answer this question (1 percent or less) were excluded from sample.
S = survey (three categories: NORC; SRC; *Washington Post*).
C = measurement context (two categories: controlled context; uncontrolled context).
n.a. = not applicable.

that since the three organizations' measurements agree in the controlled context, we can effectively rule out general organizational differences (so-called "house effects") in sampling, processing, and so forth, as an explanation for these discrepancies.

Reconciliation. While the first three experiments demonstrated the lability of general-happiness measurements, it remained to be demonstrated that any of the discrepancies that prompted our initial concerns could be accounted for by our hypothesis. This demonstration was achieved in a fourth experiment conducted for our panel in November 1980 by the Opinion Research Corporation. In this experiment, we made general-happiness measurements in contexts that replicated the low and high NORC General Social Survey measurements (see Figure 7.1).

In the negative (low) context, the general-happiness question followed three questions inquiring about the respondent's financial situation. These three questions were used in the 1973 NORC–GSS, and we had hypothesized that this context "depressed" respondents estimates of their happiness. The questions were:

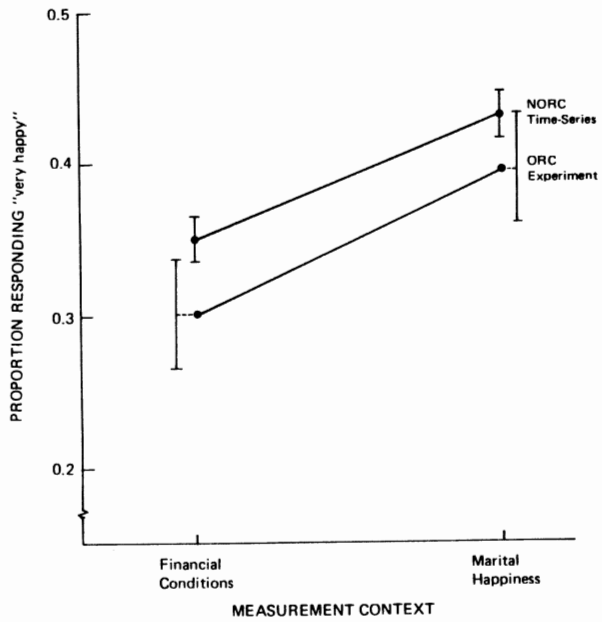
We are interested in how people are getting along financially these days. So far as you and your family are concerned, would you say that you are pretty well satisfied with your present financial situation, more or less satisfied, or not at all satisfied?

During the last few years, has your financial situation been getting better, getting worse, or has it stayed the same?

Compared with American families in general, would you say your family income is—far below average, below average, average, above average, or far above average?

(As previously mentioned [see note 15], we were led to consider the financial-happiness questions as a likely contributor to the anomalies in the NORC-GSS series because the bivariate associations between responses to the financial questions and the general-happiness items showed a significant variation between 1972 and subsequent years.) In the positive ("high") context, the general-happiness question followed the marital-happiness item.

FIGURE 7.4
Proportion of Respondents Saying They Were "Very Happy" in Response to the General-Happiness Question in Different Measurement Contexts.



NOTE: Respondents are samples of married persons; error bars demark approximately ±1 standard error around the estimates.

These two sequences replicate those used in the 1972 and 1973 NORC measurements. We hypothesized that this context variation, when experimentally induced in the 1980 experiment, would produce results that would mimic the apparent change evidenced by the NORC–GSS series between 1972 and 1973.[19]

Figure 7.4 and Table 7.3 present the relevant results. It will be seen that (for married persons) 30 percent of respondents said they were "very happy" in the negative context, and 39 percent said so in the positive context. These results are similar in direction and magnitude to those observed in the NORC-GSS time-series. When we fit formal models to the experimental and actual data (see Table 7.4), we find that a model that posits only a unitary effect of measurement context is required to account for both the experimental data and the actual data. Our hypothesis that the measurement context, and not true change, accounts for the behavior of the NORC-GSS series was thus confirmed experimentally.

Conclusion. What do we learn from this example? Recalling the general hypotheses outlined in the overview section, we observe that:

1. The concept of "happiness" is notably amorphous.
2. The happiness question involved considerable arbitrariness in the choice of a response category, for example, what is the difference between being "very happy" versus "pretty happy"?
3. The question may be one to which individuals do not give considerable thought—at least as formulated in this item (that is, Am I happy?).

TABLE 7.3
Comparison of Results of ORC Experiment and 1972–73 NORC Measurements

		Response to General-Happiness Question				
Survey	**Context**	**Very Happy**	**Pretty Happy**	**Not too Happy**	**Total**	
ORC experiment, 1980	Marital	39%	51%	11%	100%	(358)
	Financial	30	52	18	100	(393)
NORC time-series, 1972–73	Marital	43	48	9	100	(1,073)
	Financial	35	52	14	100	(1,156)

NOTE: Includes only married respondents; "don't know"/"no answer" responses excluded. Numbers may not total 100 percent due to rounding.

TABLE 7.4
Fit of Alternative Models

Model	Marginals Constrained[a]	L^2	d.f.	p
1. No effects	{CS} {H}	39.9	6	.0005
2. House-year effect only	{CS} {HS}	31.5	4	.0002
3. Context effect only	{CS} {HC}	8.6	4	.07
4. Context and house-year effects	{CS} {HC} {HS}	0.4	2	> .5
5. Interaction effects	{CSH}	0.0	0	—

NOTE: Models were fit using procedures developed by Goodman (1971). L^2 values are likelihood ratio chi-square statistics.
[a]Variables included in this analysis are:
H: Response to happiness question (three categories: very happy; pretty happy; not too happy). Respondents who did not answer this question were excluded from sample.
S: Survey (two categories: NORC-General Social Surveys: 1972, 1973; ORC survey experiment, 1980. Note that Survey house is inextricably confounded with time, i.e., NORC = 1972–73 while ORC = 1980.
C: Measurement Context (two categories: (1) financial conditions context; (2) marital happiness context.)

(We might add that happiness is not a state that is likely to be crystallized by public discussion.)

EXAMPLE II: DISAGREEMENTS ABOUT SCIENCE

Our next two examples (II and III) involve the measurement of public attitudes toward science and technology. These measurements were made in surveys commissioned by the National Science Board and conducted by the Opinion Research Corporation (ORC). The results of these surveys have been incorporated in the volumes *Science Indicators: 1972*, *Science Indicators: 1974*, and *Science Indicators: 1976*, published by the National Science Foundation.

Nonexperimental Evidence. Our interest in these surveys was first aroused by an observation[20] made during the analysis of the 1976 survey. In brief, the 1976 survey contained an anomaly that had potentially destructive implications for national science policy. It is thus a most appropriate illustration of the dangers inherent in our inadequate understanding of the error structure of the data employed as subjective social indicators.

The anomaly in the 1976 survey arose from an attempt to explore the meaning of public response to the following question about government spending for science:

> Science and Technology can be directed toward solving problems in many different areas. In which of the areas listed on this card would you *most* like to have your taxes spent for science and technology? Please read me the numbers. [*Card*: 1. Reducing and controlling pollution; 2. Finding better birth control methods; 3. Weather control and prediction; 4. Space Exploration; 5. Improving health care; 6. Developing/Improving weapons for national defense; 7. Developing faster and safer public transportation for travel within and between cities; 8. Discovering new basic knowledge about man and nature; 9. Reducing crime; 10. Improving the safety of automobiles; 11. Finding new methods for preventing and treating drug addiction; 12. Improving education; 13. Developing/Improving methods of producing food.]
>
> Please tell me the areas you would *least* like to have your taxes spent for science and technology. Again, please read me the numbers.

Data from the 1972 and 1974 Science Indicators surveys revealed that the public appeared to give relatively strong endorsement to funding science in order to reduce crime (59 percent in 1972 and 58 percent in 1974), to fight drug addiction (51 percent and 48 percent), and to improve education (41 percent and 48 percent), and relatively weak support to science spending for such purposes as the development of faster and safer mass transportation (23 percent and 26 percent), and discovering new basic knowledge (19 percent and 21 percent).

This ordering of public priorities contradicts many scientists' notions of where research could be useful, and it prompted an explicit study of this matter. In 1976 the Science Indicators survey added a question asking in what areas science and technology could (1) make a major contribution, and in what areas it could (2) make little or no contribution.[21] The same list of problem areas was used; these questions immediately preceded the questions on spending.

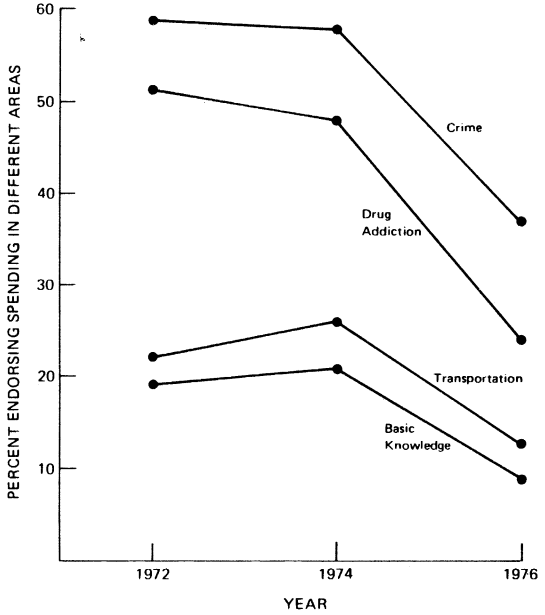
Surprisingly, neither an analysis of the relationship between the perceived usefulness of science and public endorsement of spending, nor the spending time-series appears in the 1976 report of the National Science Board. Instead, a footnote (National Science Board, 1977:180) observes that alterations in the content of the questions preceding the spending question precluded a valid comparison of the 1976 estimates to those obtained in previous years. Upon reviewing these estimates, the authors' reticence becomes very understandable. (See Figure 7.5 for plots of the more dramatic results.)

Table 7.5 presents all estimates derived from responses to the spending question in 1972, 1974, and 1976. These estimates show consistent and apparently precipitous declines in public support of spending for science and technology. In four instances, the declines (1972–76) exceeded 20 percentage points.

This "evidence" of a massive drop in public support is, however, incon-

FIGURE 7.5

Endorsement of Spending for Science and Technology in Four Areas.



SOURCE: Science Indicators surveys, 1972-76. Sample size in each year was approximately 2,100.

sistent with other independent measurements. The NORC General Social Survey has included an item on spending since 1973. The NORC series shows virtually constant levels of public support for spending in related areas between 1973 and 1976 (see Table 7.6).

Because of the changes in the content of the survey questionnaire (and, perhaps, fearing the impact of such data), the National Science Board chose *not* to plot the time-series showing the declines in their measurements of public support for science spending. Instead, they argued that

> . . . the same [spending] question was used in the 1972 and 1974 surveys, but since it was not preceded in those years by the question about the capabilities of science and technology, the results are not strictly comparable to the 1976 results. [National Science Board, 1977:180]

Our own analysis of the NORC time-series indirectly supports this argument. However, this position contradicts the prevailing wisdom among survey researchers. In their comprehensive review of past methodological research on survey measurement, Sudman and Bradburn (1974:33) concluded

TABLE 7.5
Endorsement of Spending for Science and Technology as Estimated by Science Indicator Surveys in 1972, 1974, and 1976

Area	Percentage Endorsing Spending			Change	
	1972	1974	1976	1972–76	1974–76
Improving health care	65%	69%	57%	− 8	−12
Reducing and controlling pollution	60	50	33	−27	−17
Reducing crime	59	58	37	−22	−21
Finding new methods for preventing and treating drug addiction	51	48	24	−27	−24
Improving education	41	48	33	− 8	−15
Improving the safety of automobiles	38	29	15	−23	−14
Developing faster and safer public transportation for travel within and between cities	23	26	13	−10	−13
Finding better birth control methods	20	18	10	−10	− 8
Discovering new basic knowledge about man and nature	19	21	9	−10	−12
Weather control and prediction	11	14	5	− 6	− 9
Space exploration	11	11	7	− 4	− 4
Developing or improving weapons for national defense	11	11	10	− 1	− 1
Average endorsement	34	34	21	−13	−13

NOTE: Estimates are for percentage selecting areas as ones in which they would most like to have taxes spent. See text for question wording.

that the available evidence (albeit evidence that was weak and fragmentary) indicated that

> . . . [the] position of a question [in the survey questionnaire] has by itself little biasing effect for behavioral items and a negligible effect for attitudinal items . . . [and] there do not appear to be any sizeable response effects associated with the placement of questions after related questions.

In the absence of experimental evidence upon the specific context effects postulated by the Board, it is difficult to assess the validity of these conflicting views.

Experimental Evidence. While we cannot directly test the claim made by the board, we have studied other evidence on context effects in the 1976 survey. This evidence arises because the Science Indicators surveys were amalgams consisting of several questionnaire sections sponsored by different organizations. The survey questions for the National Science Board's

TABLE 7.6
Evaluation of Government Spending Programs as Estimated by NORC General Social Surveys 1973–76

Area	Percentage Saying Spending "About Right" or "Too Little"			Change	
	1973	1974	1976	1973–76	1974–76
Improving and protecting the nation's health	95%	95%	95%	0	0
Improving and protecting the environment	92	92	90	− 2	− 2
Halting the rising crime rate	95	95	92	− 3	− 3
Dealing with drug addiction	94	93	92	− 2	− 1
Improving the nation's education system	91	91	90	− 1	− 1
Space exploration program	39	37	38	− 1	+ 1
The military, armaments, and defense	60	67	71	+11	+ 4
Average	81	81	81	0	0

NOTE: This question was not asked in the 1972 General Social Survey. Estimates are repercentaged to exclude "don't know" responses and no answers. Sample sizes in each year were approximately 1,500.
Question Wording: We are faced with many problems in this country, none of which can be solved easily or inexpensively. I'm going to name some of these problems, and for each one I'd like you to tell me whether you think we're spending too much money on it, too little money, or about the right amount. First _____ are we spending too much, too little, or about the right amount on _____?

1976 report were asked along with a melange of questions on hospitalization and medical expenditures, frequency of eating hamburgers, and the litter problem. As a partial control for context effects in these surveys, ORC administered two different versions of the survey; the order of questionnaire sections was rotated in the different versions. Also, for some multipart questions, the sequence of individual parts of a question was varied. Each version of the questionnaire was administered to (approximately) one-half of the sample.[22]

In both versions, the first item in the Science Indicators section asked respondents to assess the "prestige" or "general standing" of various occupations including "scientist" and "engineer." For the *Science Indicators* report, this question was of interest both because it was thought to be a surrogate measure of public attitudes toward science, and also because public perceptions of the prestige of scientific occupations influence the recruitment of talented young people into these professions.

For social scientists, the responses to such questions are important because they provide the basis for the well-known scalings of the socioeco-

nomic status and occupational prestige (Duncan, 1961; Hall and Jones, 1950; Treiman, 1977). These scales have been central to much recent work on social and occupational stratification (for example, Blau and Duncan, 1967; Sewell and Hauser, 1975). It is thus of considerable interest to determine whether response to this question was affected by the context variation built into the Science Indicators survey. The survey question itself read:

> I am now going to read you a list of jobs and professions. For each one I mention, please choose the statement that best gives your own personal opinion of the prestige or general standing that such a job has.

The respondent was then shown a card containing the following responses: excellent, good, average, below average, and poor, and ratings were solicited for 10 occupations (see Table 7.7).

The variation in the context and administration of the "prestige" question was twofold. First, the list of occupations used in Form A was reversed in Form B (see Table 7.7). Second, the placement of this question in the survey varied. In Form A, this question was the very first question in the survey. The interviewer began:

> [STANDARD INTRODUCTION] Hello (Respondent's name), I am (interviewer's name) conducting a study for the Caravan Surveys of Opinion Research Corporation of Princeton, New Jersey. In this interview we would like to ask your opinion on a number of different subjects.

The interviewer then read the introduction to the National Science Board's questions:

> I am now going to ask you a group of questions that come from the National Science Foundation, which is a federal agency. They are preparing a report that will discuss public attitudes toward science and technology. Your participation in this survey will be very helpful to them, but it is entirely voluntary. No records will be kept that will allow your individual reply to be associated with you.

The item on the social standing of occupations immediately followed. In Form B, the survey began with the same standard introduction, but then asked a series of 38 questions[23] on litter,[24] hamburger makers,[25] and hospitalization and medical insurance.[26] After asking these questions, the interviewer read the NSF introduction and then asked the occupational-prestige question.

We did not anticipate finding a significant divergence in the results obtained from these two forms. Our findings, however, were striking.The relevant comparisons are presented in Table 7.7. For 8 of 10 occupations,

TABLE 7.7

Variations Between Survey Forms in Percentage of Respondents Rating Occupation's Prestige as "Excellent" (rank orderings in parentheses)

Occupation	Percentage "Excellent" Form A	Form B	Discrepancy	X^2	p[a]
1. Businessman	13.4% (10)	13.4% (10)	0.0	0.0	*ns*
2. Physician	47.6 (1)	56.4 (1)	− 8.8	17 8	.005
3. Scientist	46.8 (2)	49.5 (2)	− 2.7	8.4	*ns*
4. Congressman[b]	16.1 (9)	30.4 (7)	−14.3	54.1	.0001
5. Lawyer	24.0 (6)	38.7 (3)	−14.7	41.9	.0001
6. Architect	24.6 (5)	37.5 (5)	−12.9	29.9	.0001
7. Minister	39.0 (3)	38.3 (4)	+ 0.7	9.3	*ns*
8. Engineer	25.5 (4)	34.0 (6)	− 8.5	18.9	.002
9. Banker	18.9 (7)	27.7 (8)	− 8.8	26.7	.0001
10. Accountant[c]	17.3 (8)	25.0 (9)	− 7.7	28.9	.0001

NOTES: Listing of occupations is in order used in Form A, the reverse order was used in Form B. Wording of occupational titles in table is identical to that used in questionnaire, except where noted otherwise

Chi-square tests were performed across the entire response distribution (i.e., "excellent," "good," "average," "below average," "poor," and "no opinion" response), the degrees of freedom for the tests were 5. To conserve space only, the distributions for the response category "excellent" are shown; this category accounted for a majority of the variability across forms.

[a]Computed on assumption that sampling efficiency of clustered example was 66 percent that of equivalent simple random sample See text note 22 for further discussion.

[b]Survey text: "U S Representative in Congress."

[c]Survey text: "Accountant for a large business."

prestige ratings are lower when the question is asked in the questionnaire's Form A. For 7 of the 10 occupations, this difference is 5 or more percentage points, and in 4 cases, it exceeds 10 percentage points. The sole exceptions to this general pattern occur for "Businessmen" and "Ministers"; there, the discrepancies are of trivial magnitude (0.0 percent and +0.7 percent).

Clearly, responses to this question were not identical in the two forms of the survey. Why this happened is unclear. One might speculate that survey respondents have an initial set against the use of extreme response catego-

ries (for example, excellent). This bias may diminish with practice in responding to survey questions. There is, however, some experimental evidence suggesting a modest trend in the opposite direction (Kraut, Wolfson, and Rothenberg, 1975). Alternatively, one might speculate that sequencing banal questions about beverage containers, litter, and hamburger makers immediately before questions about acute medical problems and experience with doctors and hospitals created a pro-science and pro-professional evaluation bias. The latter speculation may be plausible, especially since respondents were told that they were being asked to evaluate these professional occupations for "the National Science Foundation, which is a federal agency . . . [which is] preparing a report on public attitudes toward science and technology."

One could, of course, speculate endlessly about the causes of the observed anomaly. It is not our intention to interpret this context artifact; indeed, an interpretation would not be possible given the data at hand. Rather, we wish to know what these results tell us about our initial hypotheses. In this regard, we observe that:

1. The "prestige or general standing" of an occupation may not be a well-defined concept (particularly if the range of evaluated occupations is narrowly restricted, as in the present case).
2. The question requires an arbitrary choice between response categories (excellent vs. good vs. average vs. below average vs. poor).
3. The referents are imprecise, for example, does the "job" of businessman refer to the local grocer or the president of General Motors.
4. The *formal* rating task is one to which respondents probably give little thought in their everyday lives (although there is good reason to believe that informal ratings of jobs and occupations may be a constant and consequential part of the everyday life of most employed people).

On first impression, the foregoing results appear to be in conflict with the results of recent attempts to analyze and integrate occupational prestige data from surveys using disparate measurement methods (see, for example, Treiman's [1977] synthesis of measurements from 60 nations). On closer analysis, however, there is no inherent contradiction. Typically, attempts to integrate results from dissimilar studies of occupational prestige have relied upon the interstudy correlations as indices of agreement between the measurements. Since the prestige scales do not have substantively meaningful zero points, a general elevation or depression of scores in one study is not of concern, and it does not, in turn, affect the correlation coefficients.

If we, too, ignore the marginals and concentrate on the rankings of this restricted range of occupations, we also obtain a tolerable level of correlation between the two questionnaire forms (Spearman's rank correlation $\rho = +0.88$). Thus, the artifact in the present measurements could be finessed

by avoiding undue concentration on the univariate response distributions. Unfortunately, it is the marginal distributions that have usually been reported in the *Science Indicators* publications.

EXAMPLE III: DISAGREEMENTS ABOUT PATTERNS OF ASSOCIATION

Our next examples concern a different sort of disagreement. In the preceding sections of this chapter, we have been concerned with whether or not the univariate response distributions obtained from different surveys were comparable, for example, did two surveys provide consistent estimates of the level of public support for spending on science and technology. In the present section, we are concerned with whether patterns of association between variables measured in different surveys can vary systematically. We wish to know whether we would come to the same conclusion about the association between education, for example, and a given attitude—regardless of the survey contexts in which the measurements were made.

In this area, the prevailing wisdom is that even with major wording changes, (not to mention context) the multivariate distribution of variables will generally be undisturbed, even though the marginal (univariate) distributions can vary substantially. As one recent review noted:

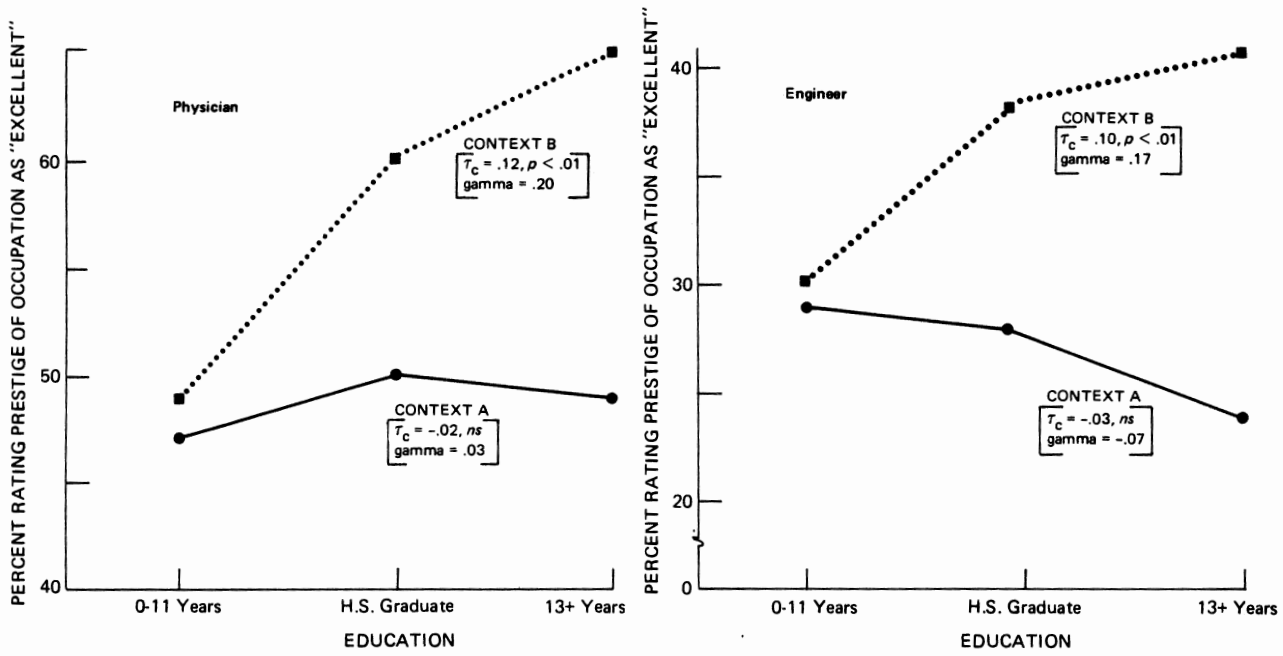
> The solution to this problem [of fluctuations in univariate distributions arising from changes in wording] advocated by . . . experienced survey investigators is to ignore single variable attitudinal results and concentrate on relationships. The assumption seems to be that single variable distributions vary for reasons that are artifactual, frivolous, or even quite meaningless, but that the ordering of respondents on items—and therefore associations among items—are largely immune to this problem. [Schuman and Duncan, 1974:234]

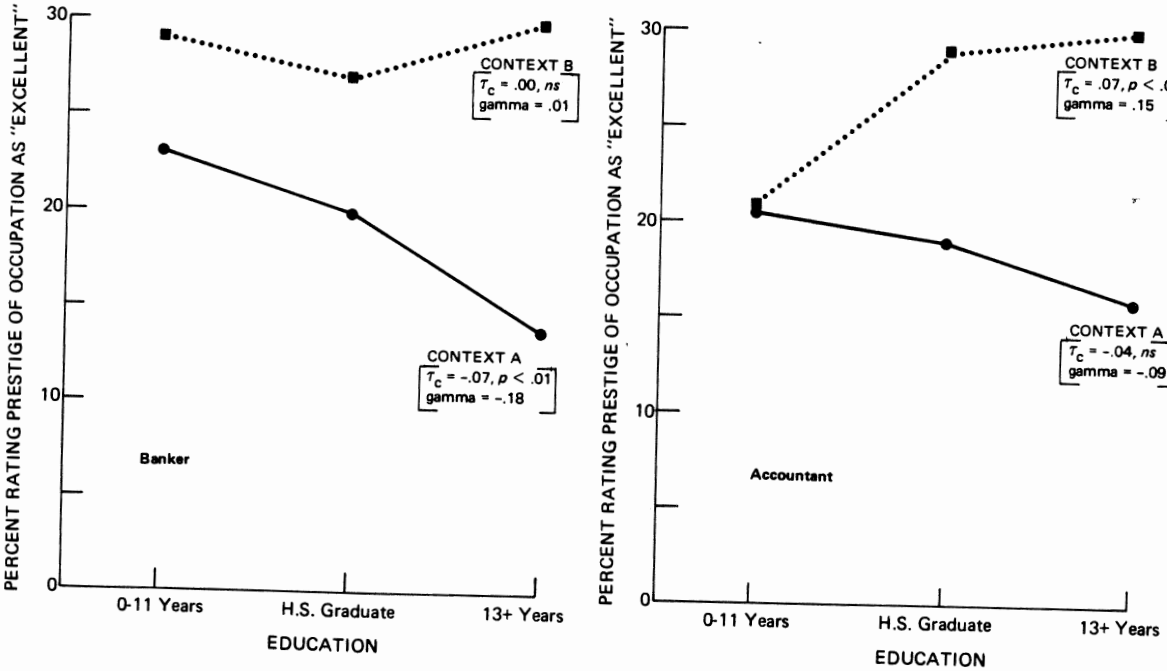
This point of view has recently been questioned, and some investigators have argued for caution in making the assumption that multivariate patterns of responses will be unaffected (see, for example, Schuman, 1974; Schuman and Duncan, 1974; Schuman and Presser, 1977; Duncan and Schuman, 1980.)

Some further information on this question can be gleaned from analysis of the *Science Indicators* data on occupations. In Figure 7.6 we plot responses to four items from the occupational prestige question by the educational level of the respondents. It will be seen from these results that the context effects observed in Table 7.7 are most pronounced for the highly educated. This, in turn, causes the bivariate patterns of association between respondents' educational level and their occupational ratings to vary systematically between Form A and Form B. Using ordinal measures of association, we observe modest (median gamma = 0.16) and generally significant positive correlations in Form A between educational level and the likelihood of

FIGURE 7.6

Relationship between Ratings of Occupational Prestige and Respondents' Educational Level for Two Forms of Questionnaire.





SOURCE: Science Indicators Survey, 1976.

rating the prestige of these occupations as "excellent." In contrast, the four correlations between prestige ratings and education are modestly negative in Form B of the questionnaire; in one instance (banker) this association is significantly negative (gamma = -0.18).[27] Thus, the conclusions one would reach about the relationship of respondents' education and their evaluation of occupations depend upon the survey context in which the questions were asked. Clearly, for these data the assumption that measurement artifacts are restricted to the univariate response distributions is unwarranted.

EXAMPLE IV: INCOMPLETE EXPLANATIONS: CONTEXT EFFECTS IN THE MEASUREMENT OF PUBLIC CONFIDENCE

In an earlier paper (Turner and Krauss, 1978), large and persistent discrepancies between Harris and NORC time-series on public confidence in the leaders of national institutions were analyzed. In that analysis, a variety of explanations for the discrepancies were investigated and discarded. These explanations included sampling variability, nonrepresentativeness of samples, the untoward effects of quota sampling, and temporal variation in public attitudes. It was concluded that the discrepancies between the Harris and NORC series, in their estimates of the level of and trends across time in public confidence, arose from the effects of large nonsampling errors in those series.

It was speculated that such nonsampling errors might arise, in part, because the questions used to measure public confidence were embedded in rather different survey contexts. These contexts varied both across survey organizations and within organizations across time. Particular attention was drawn to two instances of such contextual variation:

1. In 1976 the Harris questions on public confidence followed a series of negatively worded questions designed to measure political alienation.
2. The order in which particular institutions appeared in the confidence question varied, and in some years particular institutions were presented along with partial repetitions of the question.

It was hypothesized, in particular, that variations in use of the "alienation context" depressed the general level of confidence found by Harris in 1976, and that variations in use of a "people running" prefix accounted for the erratic behavior of NORC's estimates of confidence in organized religion. The latter effect was thought to occur because in some NORC surveys people responded to the prompt "How about the people running organized religion?" while in other surveys they were merely prompted with "organized religion."

To provide some experimental evidence upon the first hypothesis, NORC incorporated an experimental manipulation of survey context in the 1978 General Social Survey. Six "alienation" items were presented either

immediately before or immediately after the confidence question. These items read:

> Now I want to read you some things some people have told us they have felt from time to time. Do you tend to feel or not . . .
> (1) The people running the country don't really care what happens to you.
> (2) The rich get richer and the poor get poorer.
> (3) What you think doesn't count much anymore.
> (4) You're left out of things going on around you.
> (5) Most people try to take advantage of people like yourself.
> (6) The people in Washington, D.C. are out of touch with the rest of the country.

It was hypothesized that exposure to these negatively worded alienation items would depress respondents' tendency to report a "great deal" of confidence in the leaders of the various institutions. The effects of this experimental manipulation are shown in Table 7.8.

What can one conclude from these results? Certain things seem clear. First, there is evidence that this variation in context *did* produce some significant variations in estimates of the proportion of the population having "a great deal" of confidence. In particular, for the institution that immediately followed the alienation items (major companies), the difference between contexts is −7.4 percentage points. Smaller but still reliable differences were also found for two other institutions (Press: +4.8 percent; and Scientific Community: −5.2 percent). Curiously, while the alienation items generally reduced the frequency of the "great deal of confidence" response, a reverse effect was found for the press. When measured after a series of items focusing upon political alienation, measurements of confidence in the press rose.

A second conclusion we would draw from these results is that context, of the sort manipulated in this experiment, could provide only a partial explanation for the discrepancies observed between the 1976 Harris and NORC estimates. In that year, discrepancies of up to 16 percentage points were observed. In no instance did the experimental manipulation produce discrepancies of this magnitude. (We should note, however, that the NORC experimental manipulation did not fully duplicate the alienation context of the 1976 Harris survey;[28] this does introduce some uncertainty into comparisons of these experimental results to the actual survey measurements made in 1976.)

Patterns of Association. Preliminary examination of these data also revealed that the context manipulation had some significant effects upon the pattern of association between confidence and other variables. Using the three confidence items that showed significant shifts in their univariate dis-

TABLE 7.8
Effects of Experimental Manipulation of Question Context upon the Likelihood Respondents Would Express a "Great Deal of Confidence"

Institution[a]	Proportion Expressing a "Great Deal of Confidence"		Difference	X^2	p
	Neutral Context	Alienation Context			
Major companies	.264	.190	−.074	11.1	.0008
Organized religion	.329	.309	−.020	0.6	*ns*
Education	.294	.284	−.010	0.1	*ns*
Executive branch of the federal government	.126	.133	+.007	0.1	*ns*
Organized labor	.114	.117	+.003	0.0	*ns*
Press	.180	.228	+.048	5.0	.025
Medicine	.472	.456	−.016	0.3	*ns*
TV	.141	.139	−.002	0.0	*ns*
U.S. Supreme Court	.303	.285	−.015	0.5	*ns*
Scientific community	.421	.369	−.052	3.8	.05
Congress	.130	.136	+.006	0.1	*ns*
Military	.314	.299	−.015	0.3	*ns*
Banks and financial institutions	.351	.317	−.034	1.8	.18

NOTES: This analysis focuses attention upon the "great deal of confidence" category in accord with common reporting practices (e.g., *The Harris Survey*, Dec. 6, 1973; Sept. 30, 1974; Oct. 6, 1975; March 22, 1976; March 14, 1977; Jan. 5, 1978). We have eliminated missing data ("don't know," no answer, etc.) from the response distributions for each item.
[a]Institutions were presented to respondents in the same order as shown in table.
[b]Chi-square statistics have 1 degree of freedom and are corrected for continuity (Yates correction). Given that assignment to experimental conditions was fully random, the analysis treats the respondents (*N*s were approximately 1,500) as a universe and tests the hypothesis that the distribution of responses is independent of experimental condition.
Question: "I am going to name some institutions in this country. As far as the *people* running them are concerned, would you say you have a great deal of confidence, only some confidence, or hardly any confidence at all in them."

tributions, we examined the relationship of confidence to alienation and to respondents' educational level in order to determine whether there were significant context effects upon the multivariate response distributions. We

found little evidence of such effects for education. Confidence in the press and major companies showed no significant association with education in either form of the questionnaire, while confidence in the scientific community had a virtually identical association with educational level (gamma = 0.30) in the two questionnaire forms.

Subsequently we examined the association between these confidence measures and three alienation items (1, 2, and 6) we thought to be most related to confidence in national institutions. Using log-linear techniques (Goodman, 1971, 1972) to model the response distribution of Alienation (A) by Confidence (C) by Questionnaire Context (Q), we found some evidence of context effects upon the multivariate distributions. In particular, using a model that was maximally constrained to fit the observed patterns of response but which excluded the three-way interaction term {CAQ}, we could not obtain an adequate fit to the data in two of nine instances ($p < .05$), and, we obtained a rather poor fit ($p < .20$) in two further instances. Table 7.9 provides details of these analyses. We report tests only for the maximally constrained noninteractive model ({CA} {CF} {FA}), since this is the appropriate comparison for testing the null hypothesis of no context effect upon the patterns of association.

Figure 7.7 presents the actual data for the instance in which this multivariate context effect is strongest. As this figure shows, we find a considerably stronger inverse association between alienation responses ("The rich get richer and the poor get poorer") and confidence in the people running "major companies" when the alienation item follows, rather than precedes, the confidence question (gamma = 0.62 vs. 0.39). An examination of the gamma coefficients shown in Table 7.9 reveals that this particular relationship holds true in seven of the nine other comparisons. The two exceptions involve reversals of trivial magnitude.

Conclusion. Although the experiment imperfectly replicated the actual context variation, we do nonetheless observe some significant effects upon both the univariate and multivariate response distributions. These effects, however, are neither so pervasive nor so overwhelming in magnitude as to provide a complete explanation for the discrepancies observed between the Harris and NORC confidence series. Clearly, many aspects of the behavior of these series remain to be understood, and other sources of nonsampling variation in these measurements will have to be investigated.

EXAMPLE V: AGREEMENTS ABOUT FERTILITY EXPECTATIONS

Lest the reader be misled by the preceding examples, we conclude by noting that all survey measurements of subjective phenomena are not equally vulnerable to artifactual biases. With this in mind, let us consider alternative estimates of the fertility expectations of American women.

TABLE 7.9

Test for Context Effects upon Patterns of Association Between Confidence and Alienation

Alienation Item	Confidence Item	Ordinal Association[c] (gamma) Form X	Ordinal Association[c] (gamma) Form Y	Log-Linear Interaction Test[b] L^2	Log-Linear Interaction Test[b] p
The people running the country don't really care what happens to you.	Major companies	.31	.32	0.0	*ns*
	Press	.16	.15	0.0	*ns*
	Science[a]	.18	.30	1.7	.20
The rich get richer and the poor get poorer.	Major companies	.39	.62	7.4	.01
	Press	.09	.13	0.0	*ns*
	Science[a]	.24	.22	0.1	*ns*
The people in Washington, D.C., are out of touch with the rest of the country.	Major companies	.21	.39	1.8	.18
	Press	.14	.15	0.9	*ns*
	Science[a]	.07	.27	4.0	.05

NOTES: Model fit to response distribution for Alienation (A) by Confidence (C) by Questionnaire context (Q) is maximally constrained nonsaturated model. In Goodman's notation, it is {CQ} {CA} {QA}. Failure to fit a model of this type to the data indicates an interaction, i.e., that the pattern of association between the variables was not independent of questionnaire context.

[a]Item read "Scientific Community."

[b]Test has 1 degree of freedom.

[c]Form X of the questionnaire presented the confidence questions prior to the alienation items; Form Y presented them in reverse sequence.

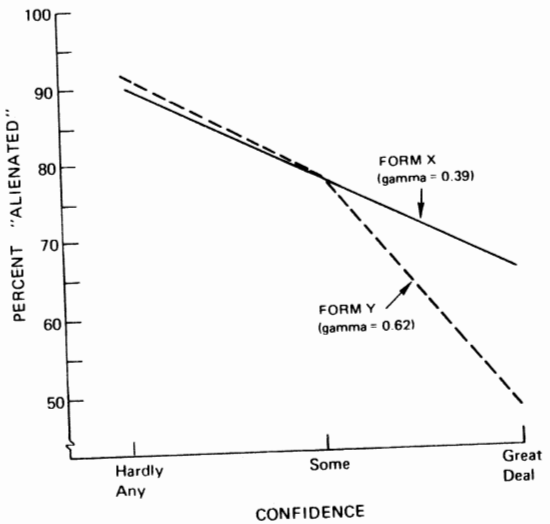
The U.S. Bureau of the Census conducts annual surveys of the birth expectations of American women. The data from such surveys are thought to be potentially useful in predicting fluctuations in the birth rate.[29] Clearly, the phenomenon being measured in such surveys is subjective. The "intention" or "expectation" of future pregnancy is not a datum subject to external verification. One must rely solely upon respondents' assessments of their own expectations or intentions.

Estimates from the Bureau of the Census's Current Population Survey (CPS) are shown in Figure 7.8 together with estimates derived from a related question on birth expectations asked in the 1972, 1975, 1976, and 1977 NORC General Social Survey. It should be noted that the latter estimates are based on modest-sized samples; on the average, there were fewer than 250 married women aged 18 to 39 in the GSS samples (versus approximately 4,000 for CPS samples). Thus, the standard errors for the GSS estimates are quite large (circa 4 percent).

Comparing the two sets of data we find that estimates of fertility expectations derived from NORC's General Social Survey are consistent with those derived from the Bureau of the Census's Current Population Surveys. In

FIGURE 7.7

Relationship of Confidence in Major Companies to Alienation Response for Two Forms of Questionnaire.



NOTE: In Form X, the confidence question precedes the alienation items; in Form Y, the alienation items precede the confidence question. The alienation item was: "The rich get richer and the poor get poorer."

only one instance (of eight) does the NORC-GSS estimate differ by more than 2 standard errors from the Census estimate.

The Lesson. The estimates of fertility expectations presented in Figure 7.8 differed in their measurement in several ways. The content of the questionnaires used to derive the estimates varied, the organizations conducting the surveys were different, and even the wording of the questions varied slightly.[30]

Because the comparison of these two series on birth expectations involved both wording and context differences, and since the measurements were made at different times of the year by different organizations, the consistency of these estimates is particularly impressive.

What lesson does this comparison teach us? In terms of our initial hypotheses, we note that:

1. The birth expectation question itself is relatively unambiguous in its meaning;
2. The response categories for the questions (for example, 0, 1, 2, 3 . . . children) have a clear meaning; and

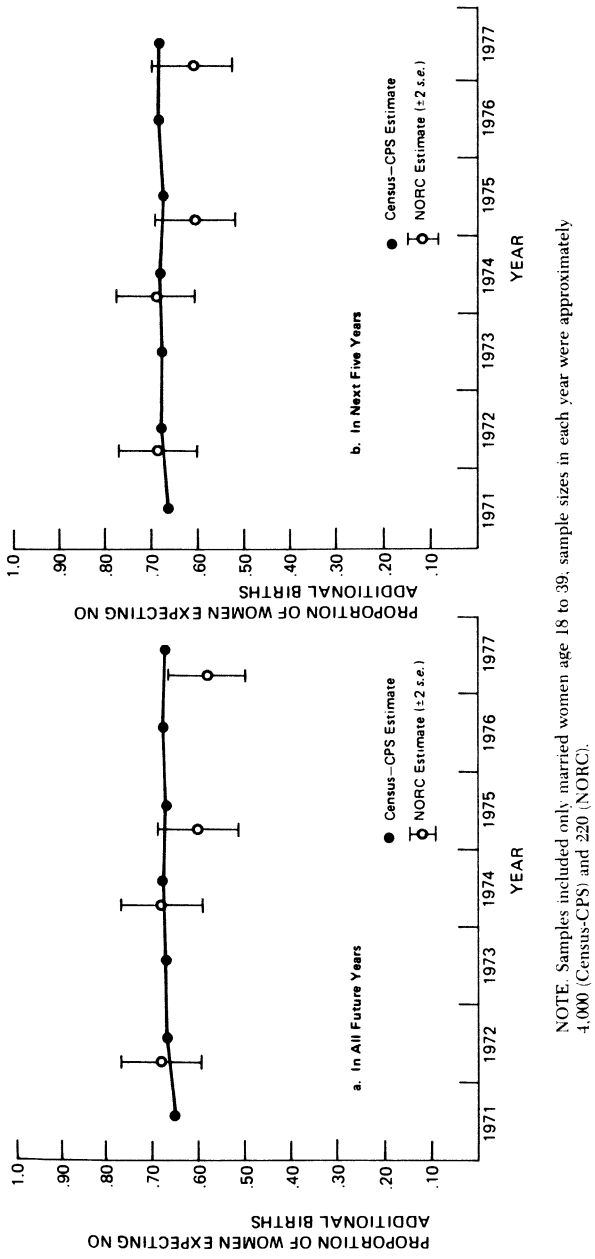


FIGURE 7.8
Estimates of Fertility Expectations of American Women: Proportion of Women Expecting No Further Children in (a) All Future Years, and (b) Next 5 Years.

NOTE. Samples included only married women age 18 to 39; sample sizes in each year were approximately 4,000 (Census-CPS) and 220 (NORC).

3. The question deals with a topic to which most respondents (that is, married women of childbearing age) should have given considerable thought. This is particularly so since attitudes toward childbearing have behavioral consequences in the everyday life of the respondents, for example, for contraceptive behaviors.

Discussion and Conclusions

In addition to providing some preliminary hypotheses, we hope that the preceding evidence will focus attention upon intersurvey comparability in the measurement of subjective phenomena. The study of disagreeable data is not an end in itself, however. To be useful, it must stimulate the difficult process of ferreting out explanations for particular anomalies and deducing general principles—where they exist.

In this regard, we believe that the foregoing evidence allows one to dispose of three commonly heard apologia for inconsistencies in survey estimates of subjective phenomena. First, because of the range of examples presented here and elsewhere (for example, Cowan, Murphy, and Weiner, 1978; Gibson et al., 1978; Smith, 1978; Turner and Krauss, 1978), these discrepancies cannot be ascribed to the deficient practices of any particular survey research organization. Such a position would be both unfair and unfaithful to the observed facts. We and others have observed both discrepancies and consistencies in comparisons involving estimates made by a wide range of research organizations.

Second, both our own analyses and those recently undertaken by Duncan and Schuman (1980) and Schuman and Presser (1977) indicate that nonsampling artifacts in the survey measurement of subjective phenomena are not limited to univariate response distributions. Hence, it is not always safe to assume that analyses focusing upon multivariate patterns of association between variables will be resistant to the anomalies encountered in the analysis of univariate distributions.

Third, it appears that no single explanation is likely to be adequate to explain all the observed discrepancies. Experimental data upon artifacts in the confidence time-series suggest that context, for example, is only a partial explanation for the discrepancies in these time-series. Various other sources of error (for example, variations in fieldwork methodology, response rates, interviewer effects, and so on) need to be considered. Basic research in this area should be encouraged.

VULNERABLE INDICATORS

At the outset of this chapter, we hypothesized that indicators of some phenomena were more vulnerable to artifacts of measurement than others.

In particular, we speculated that measurement artifacts would be more likely to afflict estimates of phenomena that were amorphous in concept, that had little importance for the everyday life of respondents, that had ambiguous referents, and that involved relatively arbitrary choices between response categories.

Stated generally, the hypothesis is open to a wide range of interpretation. To make the hypothesis more concrete, we can begin by considering limiting cases. Certainly this hypothesis predicts that survey measurements of nonsubjective phenomena, such as chronological age and years of schooling, should be relatively invulnerable to artifacts of measurement (for instance, effects of variant question wordings, survey context, and so forth). This should be so, even when we rely upon subjects' self-reports. Davis (1976) has studied survey estimates of the sex, race, age, religion, and educational distribution of the population in 30 sample surveys conducted between 1952 and 1973 by SRC and Gallup. Davis reports that these estimates do, in fact, show considerable consistency with one another and with independent estimates made by NORC. (Of course, inadequacies of sample design and execution may still cause problems with demographic measurements; see Chapter 3, Section 3.2 of Volume 1.)

Similarly, expectations of childbearing—while clearly a subjective phenomena—should be less vulnerable to measurement artifacts under our hypothesis. This follows from the fact that childbearing is an unambiguous concept (although its expectation is admittedly open to intepretation); the response categories have a clear meaning; the referent is the respondent herself; and the question itself is directly relevant and has behavioral implications in the everyday life of respondents (married women of childbearing age). Our analysis of eight pairs of estimates of women's fertility expectations revealed a pattern of consistency that was within the range expected on the basis of sampling fluctuations.

The discrepancies observed in our Examples I through IV are also consistent with our preliminary hypotheses. In Table 7.10, we summarize these examples together with other comparisons made elsewhere (Turner and Krauss, 1978; Smith, 1978; Kalton, Collins, and Brook, 1978).[31]

This summary of 101 comparisons for 21 survey questions shows a rough correspondence with the typology of our preliminary hypotheses. Thus, we find relatively more significant discrepancies for the questions (1–14) at the top of the table. These questions involve measurements of rather amorphous concepts such as confidence in the people running institutions, and evaluations of occupational prestige, national spending, and contemporary driving standards. These same items also require a choice between relatively imprecise response categories, such as

1. great deal, only some, hardly any (confidence)

2. too much, about right, too little (spending)
3. excellent, above average, average, below average, poor (prestige)

In contrast, questions (15–21) about the legalization of marijuana, gun control, the death penalty, political party affiliation, and fertility yield relatively few discrepancies. The latter questions involve somewhat less amorphous topics and response categories, for example, support or nonsupport of legislation, the name of an actual political party, or an actual number of children expected.

One further difference between these two groups of questions is the (likely) salience of the topics for respondents. The latter group of questions inquire about topics that have been the subject of considerable public discussion (for example, capital punishment and the legalization of marijuana), or that are connected with specific behaviors that often have tangible behavioral components (for example, voting, party registration, and contraceptive practice). For this group of questions, only two significant discrepancies were observed in 19 comparisons. In contrast, the first group of items asked about topics which, we suspect, would not be subject to considerable public discussion as phrased in these questions (e.g., Do I have confidence in the people running organized religion? What is the prestige of accountants? Are we spending too much on science? etc.) More than one-half of the comparisons (42 of 82) involving this group of questions produced significant discrepancies.

While we believe that a case can and has been made for the general typology we hypothesized, Table 7.10 is not without its counterexamples. Voting for a woman president (Yes or No) is a concrete action and related to issues that have been the subject of considerable public discussion. Yet, in the one instance where a comparison was possible, there was a modest (4 percentage points) but reliable difference between estimates derived from independent surveys using this question. Thus, while the evidence is broadly compatible with our hypotheses, the correspondence between our typology and the available data is less than perfect.

Other recent evidence (not included in Table 7.10) seems to fit this typology of vulnerable indicators. For example, in an unpublished experiment, Duncan and Schuman found significant context-induced variations in responses to five (of seven) survey questions. These questions measured agreement with statements such as

- Public officials really care about what people like me think.
- Given enough time and money, almost all of man's important problems can be solved by science.

Respondents chose among four response categories: strongly agree, agree, disagree, strongly disagree.

TABLE 7.10

Examples of Discrepancies in Survey Measurements of Subjective Phenomena

Topic	Response Categories	Type of Comparison[a]	Number of Comparisons	"Significant" Discrepancies[b]
1. Confidence in people running national institutions	Great deal of confidence, only some confidence, hardly any confidence	Different houses[c] Context experiment[i]	27 13	18 3
2. Evaluation of amount of federal income tax	Too high, about right, too low	Context experiment[c]	1	1
3. General happiness	Very happy, pretty happy, not too happy	Different houses[d,i] Same house/different surveys[i,j]	2 2	1 1
4. Evaluation of prestige of occupations	Excellent, good, average, below average, poor	Context experiment[i]	10	7
5. Evaluation of contemporary driving standards	Lower than they used to be, higher than they used to be, about the same	Context experiment[f]	2	1
6. Evaluation of traffic noise	Becoming noisier, less noisy, or about the same	Context experiment[f]	2	1
7. Evaluation of need for nighttime truck deliveries to retail stores	Yes, no[g]	Context experiment[f]	2	1
8. Evaluation of spending on national problems	Too much, about right, too little	Different houses[h]	10	5
9. Misanthropy I: people try to take advantage of you	Try to take advantage if they got a chance, try to be fair	Different houses[h]	2	0
10. Misanthropy II: most of time people try to be helpful	Try to be helpful, mostly just looking out for themselves	Different houses[h]	2	1
11. Misanthropy III: most people can be trusted	Can be trusted, can't be too careful in dealing with people	Different houses[h]	2	1
12. Courts' treatment of criminals	Deal too harshly or not harshly enough with criminals	Different houses[h]	1	0
13. Men or women better suited for politics	Men better suited, equally suited, women better suited	Different houses[h]	1	0
14. Vote for woman for president	Yes, no	Different houses[h]	1	1
15. Children safe bicycling in local area	Yes, no	Context experiment[f]	2	0
16. Law requiring police permit to purchase gun	Favor, oppose	Different houses[h]	3	0

TABLE 7.10 (*continued*)
Examples of Discrepancies in Survey Measurements of Subjective Phenomena

Topic	Response Categories	Type of Comparison[a]	Number of Comparisons	"Significant" Discrepancies[b]
17. Support for legalization of marijuana	Yes, no	Different houses[h]	1	0
18. Political party with which one identifies self	Republican, Democrat, Independent, other	Different houses[h]	3	1
19. Favor death penalty for murder	Yes, no	Different houses[h]	3	0
20. Ideal number of children for family	0,1,2,3 . . .	Different houses[h]	1	0
21. Expected number of children	0,1,2,3 . .	Different houses[e,i]	8	1

NOTES: These examples were derived from comparisons cited in (Kalton, Collins, and Brook, 1978; Smith, 1978; Turner and Krauss, 1978; and Turner, 1981a). Only survey measurements of subjective phenomena are included; reports of behaviors (e.g., voting) and demographic estimates included in the foregoing sources were excluded. Comparisons involving variant question wordings have been excluded—except where noted otherwise.

Appendixes F and G in Volume 1 present a more comprehensive list of (nonexperimental) intersurvey comparisons. Chapter 16 (Volume 2) present a separate assessment of the reliability and validity of 513 survey measurements of presidential popularity made during 1963–1980 by the Gallup Organization, Louis Harris and Associates, CBS/*New York Times*, NBC/AP, and the Roper Organization.

[a]All comparisons involve either (1) independent estimates made by different survey research organizations (different houses); (2) experiments involving the manipulation of the context in which survey questions were presented (context experiment); or (3) estimates derived from separate surveys conducted by the same survey research organization (same house/different surveys).

[b]Discrepancies are absolute differences estimated proportions of population giving a particular response in two surveys (or two experimental conditions). Discrepancies have been terms "significant" if the observed discrepancy exceeded twice the sampling error for the difference between the proportions. Published statistics and/or standard errors from the source publications were used, if available. When data were derived from clustered samples and no published statistics were available, estimates were based on the formula for simple random samples, but the sample size was deflated ($N' = 0.66N$) to allow for sample design inefficiencies.

[c]Turner and Krauss (1978).

[d]Comparison involves a modest difference in question wording.

[e]Note that all comparisons involve a subset of GSS sample that numbered less than 250 each year.

[f]Kalton, Collins, and Brook (1978)

[g]Source does not quote explicit response categories; they are implied by text of question.

[h]Smith (1978).

[i]Turner (1981a) and present paper.

[j]Comparisons were made by pooling data from two months of NORC Continuous National Survey closest to dates of General Social Surveys.

Similarly, studies of Census Bureau estimates of crime victimization (Cowan, Murphy, and Weiner, 1978; Gibson et al., 1978) have revealed that substantial variations have been induced by differences in survey contexts. Gibson and associates (1978) report that this measurement artifact produced a relative increase of 12 percent in the reported rate of property crime and 21 percent in the reported rate of personal crime. Examination of the National Crime Survey questionnaire suggests that these results may fit within the foregoing typology. The victimization questionnaire includes, for example, the following screening question:

> Q. 42 Did anyone try to attack you in some other way? [other than attacks already mentioned]

The concept of "attempted" attack in this question is somewhat amorphous. (When, for example, does an accidental bump become a hostile blow?) Cowan, Murphy, and Weiner (1978:282) find that reported rates of "attempted assaults (without weapon)" showed significant variation across measurement contexts (Z of difference = 2.76) and that "attempted assault (with weapon)" showed a difference of borderline significance (Z = 1.83); however, Cowan, Murphy, and Weiner find that the reported rates for assaults "with injury" showed no significant variations across measurement contexts (Z = 0.17 and 0.27). This pattern of results suggests that the vulnerability of these reports to context effects decreased as the events in question became more unmistakably hostile and, we assume, more salient to the victims.[32]

AREAS FOR RESEARCH

The growing use of survey measurements of subjective phenomena in policymaking and social research make it important for us to understand the sources of variability that affect comparisons made across different surveys (and/or surveys conducted by different survey organizations). Intersurvey comparisons that rely exclusively upon sampling error computations can be dangerously misleading when the error structure of our survey estimates include large error and bias components contributed by nonsampling factors (for example, variations in fieldwork methods, interviewer training, response rates, questionnaire context, and so forth.). The preceding examples provide evidence of the untoward effects that can be induced by the nonsampling variances involved in such comparisons. A better understanding of the error structure of these measurements is needed.

A useful first step in a future research program might involve the testing of hypotheses concerning the types of indicators that are more (and less) vulnerable to such nonsampling errors. The results of such studies might be helpful in delimiting a domain of measurements that are robust. Relatively

simple research designs might involve contemporaneous measurements made by a number of survey organizations. The extent to which the distribution of such derived estimates departs from what is expected on the basis of sampling error alone can provide some insight into the relative magnitudes of the nonsampling components of variance for different types of measurement.[33] If the preliminary hypotheses we have suggested are correct, we would expect some orderliness in these results.

To operationalize our preliminary hypotheses, we might consider some of the following procedures. Smith (1979a) studied the confidence question by asking respondents in the 1978 General Social Survey to describe what they thought this question meant. Respondents' answers were often sobering; for example, 34 percent of respondents could name no specific group that they thought was part of the "scientific community," and 2.3 percent of respondents thought that this term referred to their local community.

Similar probing of respondents' interpretation of other common survey questions might be done. The variability in respondents' interpretations might be taken as a measure of the degree to which questions are amorphous in concept or ambiguous in their referents. Our hypotheses would predict that larger nonsampling errors would be found for items that have greater variability in interpretation. Parallel procedures might be used to assess the extent to which response categories require respondents to make arbitrary choices. (See Bradburn and Sudman [1979:Chapter 10] and Hakel [1968] for relevant discussions.) The frequency with which respondents select an offered "don't know" category of response might also be a useful indicator of whether a particular topic has a well-defined place in public discussion.[34]

In addition to studies of contemporaneous measurements replicated across houses, planned experiments should be considered. Where specific measurement artifacts are thought to complicate the use of past data, experiments incorporated in future surveys may allow estimation of the magnitude of such artifacts and appropriate adjustments of past measurements (see Duncan and Schuman, 1980, for an example of such procedures). These experiments themselves might be contemporaneously replicated by several survey organizations to ensure the robustness of important measurements.

In addition to research involving the collection of new data, useful insights may also be derived by reconsidering data presently available in survey archives. For example, several organizations routinely use two alternative forms of the survey questionnaire in their amalgam studies (for example, the ORC Science Indicators study). Data from these studies would provide an inexpensive basis for post hoc experimental studies. Moreover, the advent of computer-assisted telephone interviewing (see

Chapter 2 of Volume 1; and Shure and Meeker, 1978) will provide the opportunity for easier experimentation with alternative questionnaire forms and wording in future research, and the possibility for more routine estimation of the interviewer component of the nonsampling variances in our measurements.

FUNDAMENTALS AND FUTURE DIRECTIONS

We would also suggest that there is a need for a reconsideration of the assumptions that underlie the practice of survey research.[35] The most fundamental phenomena of survey research are quintessentially social psychological in character.[36] They arise from a complex interpersonal exchange, they embody the subjectivities of both interviewer and interviewee, and they present their interpreter with an analytical challenge that requires a multitude of assumptions concerning, among other things, how respondents experience the reality of the interview situation, decode the "meaning" of survey questions, and respond to the social presence of the interviewer[37] and the demand characteristics of the interview.

The burden of the observed anomalies should prompt a reconsideration of the social psychological foundations of survey research. The foregoing examples are indicative of the deficient state of our present knowledge. We doubt any instant solutions exist, but it seems clear that complacency will not suffice.

Notes

1. In accord with traditional usage, we treat the *possibility* of independent verification (corroboration) as a litmus test for classifying phenomena as subjective or nonsubjective. This position assumes the existence of a nonsubjective (that is, "objective") reality whose properties are potentially discoverable through some consensual process. (We, of course, need not make such an assumption; see, for example, the writings of Bishop Berkeley or the radical skepticism of Descartes' first meditation.)

2. Observation of a subject's behavior can only provide an indirect measure of that subject's attitudes, beliefs, and so forth. Conceptually we can distinguish between behaviors and attitudes; in practice, experience affirms that an individual's behavior need not accord with his or her beliefs. Indeed, we should bear in mind that typical measurements of subjective states involve inferences made from subjects' behavior(s), that is, their verbal behavior in response to a question. Such response behaviors are used to infer the state and structure of subjective reality. However, as behaviors, such responses are only indirectly and assumptively linked with the inferred phenomenon (the subjective state of the individual). Among others, misreporting and measurement artifacts are common pitfalls in this inference process. The general topic of attitude-behavior inconsistency has itself received considerable attention in the social psychological literature (see, for example, the review by Schuman and Johnson [1976]).

3. See, for example, the work of the evaluation and research program of the Census Bureau or the valuable series of recent studies by Norman Bradburn, Seymour Sudman and their associates (1979) on response errors for "threatening questions" (for example, measures of the incidence of bankruptcy, drunkenness, and so forth).

4. For example, to be classified as "unemployed" in the Bureau of Labor Statistics index, workers must be both out of work and "looking for work." While the definition of the criterion "looking for work" is not left to the discretion of respondents, statements that a worker has "checked with friends or relatives" during the last 4 weeks meet the criterion. Persons without paid employment who are available for work and who have checked with friends are defined as "unemployed"; those who have not checked with friends or otherwise searched for a job are excluded from the unemployment tabulation and are not counted as part of the active labor force. See National Commission on Employment and Unemployment Statistics (1979:Chapter 4) for a further discussion of the subjective components of such measurements.

5. The meaning and interpretation of questions such as "Did anyone try to attack you . . . ?" are supplied by the respondents and may differ from respondent to respondent and across the contexts in which survey measurements are made (see Martin, 1978, 1981, 1983).

6. An independent review (Caplan and Barton, 1976) of the use of the first *Social Indicators* volume (1973) concluded that there was a need for federal statistical compendiums to "go beyond objective indicators and provide subjective measures of life experience and social well being."

7. In the board's words (1975:145):

> Public attitudes affect science and technology in many ways. Public opinion sets the general environment and climate for scientific research and technological development. It is influential in determining the broad directions of research and innovation, and through the political process, the allocation of resources for these activities. In addition, public attitudes toward scientists and engineers and their efforts affect the career choices of the young by influencing their decision to enter these fields.

8. Similarly, there was a unique, albeit short-lived, attempt to provide federal policy makers with easy access to an ongoing series of national surveys (NORC's Continuous National Survey). The latter project was funded by the RANN (Research Applied to National Needs) Division of the National Science Foundation (NSF) for the use of the Departments of Agriculture, Health, Education and Welfare; Housing and Urban Development; and Transportation; the Office of Management and Budget; the President's Commission on Gambling; and the NSF. For a history of this pioneering but ill-fated program, see, Rich (1975).

9. Parallel efforts are evident in Europe (see, for example, Abrams, 1973; OECD, 1974). In addition, the United Nation's report, *Toward a System of Social and Demographic Statistics*, voiced a concern for the inclusion of measures of subjective phenomena in the national statistics of member countries (United Nations, 1975:32).

10. Turner and Krauss (1978) hypothesize that

> . . . the degree of variability in indicator estimates derived from different surveys may be a function of an identifiable dimension of the indicator questions themselves. Thus, we know that sample surveys can show remarkable consistency in their estimates of the demographic characteristics of the population and that election predictions have been quite accurate in recent years. Furthermore, attitudes that have a relatively well-defined place in

public discussions, such as attitudes toward capital punishment, fertility expectations, and political party identification, also seem to yield rather consistent estimates in different surveys. However, there appears to be emerging some evidence that the most unstable indicators are those involving questions that are the most amorphous in their meaning (e.g., What is "confidence" or "trust"?), the most imprecise in their referents (e.g., Who are the people "in charge of running organized religion"?), and that involves the most arbitrariness in the selection of a response category (e.g., We know what it means to be "for" or "against" the death penalty, but what does it mean to have "a great deal of confidence" vs. "some confidence" or to be "very happy" vs. "pretty happy"?)

11. This question was first incorporated in a 1957 survey conducted by the Survey Research Center (see Gurin, Veroff, and Feld, 1960). A similar question, varying slightly in wording, has been asked by Gallup since 1946 (see Hastings and Southwick, 1974:400).

It should be noted that we report in this example (see Figure 7.1) only the disagreement that prompted our concerns. For a review of other measurements, see Smith (1979b).

12. The meeting was hosted by the Institute for Research in the Social Sciences, University of North Carolina, and was attended by J. Davis, O. D. Duncan, E. Martin, F. Munger, R. Parke, M. Schulman, H. Schuman, T. Smith, G. Taylor, and C. Turner. We first presented these data at the meetings of this working group. Subsequently, one member of the group conducted his own study of these data. He independently reached a conclusion similar to ours concerning the likely cause of the observed discrepancies (Smith, 1979b).

13. If large month-to-month fluctuations are taking place, then yearly or biennial measurements (taken in different months) are of little value in tracking annual changes.

14. The divergence in these trends is only suggestive, since analysis of the data indicates that a model positing only effects for year and marital status provides a tolerable fit for these data. An interaction term for the context effect (year × marital status × happiness) is not strictly required.

15. In particular, we found a significant interaction of survey Year (Y: 1972 vs. 1973–78) with responses to the Happiness (H response: very happy vs. pretty happy vs. not too happy) and Financial Status (F) questions. Analyzing data for the two financial-status questions separately, we found that models incorporating all two-way interactions (that is, {HY}, {HF}, {YF}) did not provide an adequate fit, and there was a decline in the rank correlations. The first financial-status question was "We are interested in how people are getting along these days. So far as you and your family are concerned, would you say that you are pretty well satisfied with your present financial situation, more or less satisfied, or not satisfied at all?" Using responses to this question, the likelihood ratio chi-square for the model was 11.8 with 4 degrees of freedom ($p = .02$). The rank correlation (gamma) declined from +0.47 (1972) to +0.40 (1973–78).

The second question was "During the last few years, has your financial situation been getting better, getting worse, or has it stayed the same?" Using responses to this question, the likelihood ratio chi-square for the model was 11.7 with 4 degrees of freedom ($p = .02$). The rank correlation (gamma) declined from +0.34 to +0.28.

Note that since the marital-happiness question was not asked in 1972, a parallel analysis of the bivariate associations between marital and general happiness was not possible.

16. Immediately *preceding* the experimental general-happiness and marital-happiness questions, respondents were asked the questions shown below:

SRC experiment

C74. Thinking ahead *3 or 4 years*, do you think you will be able to get all the gasoline you want, or will there be problems getting all the gas you want 3 or 4 years from now?

C75. Do you think that the price of gasoline will go up during the next 12 months, or will gasoline prices stay about the same as they are now?

IF answer to C75 was "go up":

C75a. About how many cents per gallon do you think gasoline prices may increase during the next 12 months compared to now?

C75b. If you were going to buy a new car or truck in the next year, would gasoline prices increasing by _____ (CENTS FROM C75a) per gallon affect your decision in any way?

C75c. How would it affect your decision to buy a new car or truck?

C76. One government proposal has been to ration gasoline at 2 gallons per vehicle per day. You could obtain additional gasoline but you would pay a higher price for it. Would this rationing proposal affect your plans to buy or lease a car or truck in the next 12 months?

IF Answer to C76 was "would affect plans":

C76a. In what ways would this affect your plans to buy or lease a car or truck in the next 12 months?

C77. There has been a lot of talk about our nation's ability to pay for the oil that we import from foreign countries. Do you believe this is a serious problem, or is it not very serious?

C77a. Why do you say so?

Now we would like to ask a few questions about you (and your family living there).

D2. Are you currently married, separated, divorced, widowed, or have you never been married?

Washington Post Poll experiment

E. In what year were you born?

F. What religion were you brought up in—Protestant, Catholic, Jewish, or something else?

G. How would you describe the place where you live: Is it a large city, a suburb of a large city, a small town, or a rural area?

H. We'd like to know if the chief wage earner in your household is working now, unemployed, retired, or what?

I. *IF Working*: What sort of work does he/she do?
IF Unemployed: What sort of work did that wage earner do on his/her regular job?

J. What is your race? Are you white, black, hispanic or what?

K. If you added together the yearly income before taxes of all the members of your household last year in 1978, would the total be under $8,000; between $8,000 and $12,000; between $12,000 and $20,000; between $20,000 and $30,000; between $30,000 and $50,000, or would the total be over $50,000 [IF UNCERTAIN, ASK:] What would be your best guess?

Now I would like to ask you one or two more qestions concerning your general feelings about life. . .

NORC experiment

For each area of life I am going to name, tell me the number that shows how much *satisfaction* you get from that area. (READ ITEMS A–E. CIRCLE ONE CODE FOR EACH.)

A. The city or place you live in.

B. Your non-working activities—hobbies and so on.
C. Your family life.
D. Your friendships.
E. Your health and physical condition.

17. The data that we analyze in the following pages have been restricted to cases in which respondents reported that they had a telephone in their residence, in order to ensure comparability between the NORC sample and the samples used in the SRC and *Washington Post* experiments. (Approximately 3 percent of the NORC sample said they had a telephone but refused to supply a telephone number and thus the interviewer did not ascertain the exact location of the phone. On the assumption that most of these telephones were probably in the respondents' residences, these respondents were maintained in the sample we used for our analyses.) All data used in our analyses of these experiments are unweighted. Due to an error in the preliminary processing of the SRC data, six eligible respondents were eliminated from our tabulations. This error changes none of the SRC marginals by more than 1 percent.

Respondents were randomly assigned to experimental conditions in all experiments. In the SRC and *Washington Post* experiments, sample assignments were divided 50–50 between the two conditions; in the NORC experiment the division was 66–33. Selection of individual respondents within households was done by random selection from among all members of the households in the NORC and SRC surveys; in the *Washington Post* survey, selection was done from among those who were at home at the time of the initial contact.

18. The tabulations from the SRC experiment were independently confirmed by Howard Schuman, who also arranged for a subsequent hand-check of several questionnaires to ensure that the experimental conditions were not systematically miscoded or erroneously transcribed.

19. The experimental questions were introduced with the statement "These questions deal with several different topics—" followed by the inquiry:

X1. Are you currently married, separated, divorced, widowed, or have you never been married?

In the NORC-1972 replication, the respondents were then asked:

Y2. We are interested in how people are getting along financially these days. So far as you and your family are concerned, would you say that you are pretty well satisfied with your present financial situation, more or less satisfied, or not satisfied at all? (IF RESPONSE IS ANYTHING OTHER THAN THE RESPONSE CATEGORIES GIVEN, PROBE:) In general, (THEN REPEAT QUESTION).

Y3. During the last few years, has your financial situation been getting better, getting worse, or has it stayed the same? (IF RESPONSE IS ANYTHING OTHER THAN THE RESPONSE CATEGORIES GIVEN, PROBE:) In general, (THEN REPEAT QUESTION).

Y4. Compared with American families in general, would you say your family income is—far below average, below average, average, above average, or far above average? (IF THERE IS ANY DIFFICULTY ANSWERING, SAY:) Just your best guess.

Y5. Taken altogether, how would you say things are these days—would you say that you are very happy, pretty happy, or not too happy? (IF UNDECIDED, PROBE:) In general, how do you feel?

In the NORC-1973 replication, respondents were asked the marital status question X1, and then:

(IF "MARRIED" ON Q. X1, ASK):

X2. Taking things altogether, how would you describe your marriage? Would you say that your marriage is very happy, pretty happy, or not too happy? (IF UNDECIDED, PROBE:) In general, how do you feel?

X3. Taken altogether, how would you say things are these days—would you say that you are very happy, pretty happy, or not too happy? (IF UNDECIDED, PROBE:) In general, how do you feel?

In addition to these conditions, approximately one-third of respondents were asked questions in a format designed to mimic the gas shortage context of the August 1979 SRC measurements. These respondents were questioned as follows:

These questions deal with several different topics:

Z1. Do you think that the price of gasoline will go up during the next 12 months, or will gasoline prices stay about the same as they are now?

Z2. There has been a lot of talk about our nation's ability to pay for the oil that we import from foreign countries. Do you believe this is a serious problem, or is it not very serious? Why do you say so?

Z3. Are you currently married, separated, divorced, widowed, or have you never been married?

Z4. Taken altogether, how would you say things are these days—would you say that you are very happy, pretty happy, or not too happpy? (IF UNDECIDED, PROBE:) In general, how do you feel?

The results of this mimicry of the gas shortage context did not, however, provide results comparable to those obtained by the August 1979 SRC experiment. We suspect that this result may be due to both the incomplete context replication which was done and to the general easing of the national gasoline shortage in the intervening period.

20. Communicated to us by Dr Donald Buzzelli.

21. The question read: "This card lists a number of areas in which there are problems. In your view, in which of these areas could science and technology make a major contribution toward solving the problems? Please read me the numbers." [Card: same as spending question.] "In your view, in which areas could science and technology make little or no contribution? Please read me the numbers."

22. Selection of respondents to receive the different versions was performed as follows: The survey itself used three separately generated national samples. In Sample 1: 60 sampling points received Form A and 60 received Form B. For Sample 2, 30 sampling points received Form A and 30 sampling points received Form B, and for Sample 3, 50 sampling points received Form A and 50 received Form B. All forms were randomly assigned to sampling points.

Randomization by sampling points rather than individual respondents causes some difficulties. In particular, we know that clustering of sampling points results in a departure of sample efficiency from that of SRS designs; the extent of this departure is impossible to compute given the presently available data. In the analysis of the confidence items, Turner and Krauss (1978: 461-462) found that the use of clusters of 15 as the unit for analysis produced a deflation of the effective sampling size from 1,500 to 1,000 (using the level of intracluster correlation as a deflator [Blalock, 1972:Chapter 20; Kish, 1965]). Whether that result holds here is unknown; however, we would point out that analysis of the demographic characteristics revealed no significant differences between the samples receiving the different versions of the questionnaire. (Note: Computations assume effective sample size to be deflated by 0.66—that is, N = 2,000 becomes N = 1,333.)

Variable	X^2	d.f.	p
Sex	0.0	1	*ns*
Education	4.9	9	*ns*
Age	8.5	8	.40
Income	13.2	11	.27
Religion	4.6	4	.35
Marital status	0.8	4	*ns*

23. This survey, the ORC Caravan, consisted of eight parts—each funded by different organizations. The Opinion Research Corporation treats each section of the questionnaire as the confidential property of the sponsoring organization. Thus, ORC has not been able to make available to us the actual questions used in each section. However, Dean Behrend and his staff have made available a summary of the content of each section and have been helpful in answering questions about survey administration. The following notes describing the content of survey sections are derived from summaries prepared by ORC staff.

24. (1) Seriousness of litter problem in United States; (2) who is responsible for problem; (3) degree of activity of eight organizations in fighting litter problem; (4) awareness and sponsor identification of antilitter advertising; (5) use of one-way or returnable packaging in purchase of beverages.

25. (1) Frequency of serving hamburgers in household; (2) how many cooked at one time; (3) ownership and usage of hamburger makers (electrical appliance); (4) purchase intention and brand intention for a hamburger maker.

26. (1) A series of questions on hospitalization insurance, family members covered, and so forth; (2) proportion of expenses paid by such coverages; (3) hospitalization incidence last year; (4) estimated cost of hospitalization.

27. Categorical analyses of the multivariate response distributions yields similar but weaker results. Treating the three categories of education (E) as an unordered classification, the patterns of response (R) to different forms (F) were only poorly fit by models that excluded the three-way interaction term {ERF}. In particular, models fitting only the marginals for the {ER} {FR} {EF} distributions produced the following fits to the data:

Occupation	L^2	P
Engineer	5.4	.07
Physician	4.3	.11
Accountant	5.0	.08
Banker	3.9	.14

In each case the degrees of freedom associated with the test of fit were 2; computations assumed that clustering effects diminished the effective sample size by one-third.

28. In particular, two questions that were interspersed between the alienation and confidence questions in the 1976 Harris survey were omitted from NORC's experimental context manipulation. These questions were (1) "Compared to 10 years ago, do you feel the quality of

life in America has improved, grown worse, or stayed the same?" (2) "Compared to 10 years ago, do you feel the leadership inside and outside of government has become better, worse, or stayed the same?" The questions about confidence in "the people running" national institutions followed these items in the original Harris survey.

29. Longitudinal studies indicate that these measures do predict subsequent fertility; see, for example, Wilson and Bumpass (1973); Freedman, Hermalin, and Chang (1975); Goldberg, Sharp, and Freedman. (1959).

30. The text of birth-expectations questions used by Census and NORC between 1972 and 1977, was as follows:

Organization	Years	Text
Census	1972–74	1. Do you expect to have any (more) children? 2. How many (more) do you expect to have? 3. How many (more) do you expect to have in the next 5 years?
	1975–77	1. Looking ahead, do you expect to have any (more) children? 2. How many (more) do you expect to have? 3. How many (more) do you expect to have in the next 5 years?
NORC	1972–77	1. Do you expect to have any (more) children? 2. How many (more)? 3. How many (more) in the next 5 years?

31. We have included only recent studies because the earlier literature was reviewed by Sudman and Bradburn (1974); their conclusions in some areas, however, are somewhat different from those that have emerged from later work.

32. A somewhat similar pattern was found for reports of thefts. Reports of larcenies of under $50, for example, evidenced significant context-induced variations (Z = 3.07), but those involving larger amounts of money and reports of auto thefts did not vary significantly between measurement contexts (Z = 1.20, larceny of $50 or more; 0.15, theft of car [Cowan, Murphy, and Weiner 1978:282).

33. In this regard, the procedures developed to assess analytical measurements made by different laboratories in the physical and chemical sciences may offer useful guidance (Youden, 1975; Steiner, 1975).

34. See Schuman and Presser (1978) for a thought-provoking discussion of the effects of offering a "don't know" alternative upon multivariate response patterns.

35. In this regard, we applaud those independent initiatives that have recently emerged (for example, Nisbett and Wilson, 1977; Wilson and Nisbett, 1978; Fischhoff, Slovic, and Lichtenstein, 1979).

36. I am grateful to O. D. Duncan for reminding me that just as the fundamental phenomena of survey research are psychological, so too they are fundamentally sociological. In his words, the topics we commonly survey largely

> . . . have to do with political categories, economic sectors, institutional arrangements, or other social objects, entities, or systems. Hence, if most of the content of polls is subjective and hence psychological in one aspect, it is also true that most of the content is sociological (including, for simplicity, economic, political, cultural, etc. content under this heading). The statements are not contradictory, of course, but there is reason to suspect that survey work can go seriously astray when the productive tension between the two ways of looking at survey content is disregarded. Specifically, the survey researcher, whether interested in subjective or objective variables, does well to keep in mind the fact that interviewer and respondent are both fallible with respect to their supposed functions in making observations and reports. It appears that we have come to suppose an extraordinary proficiency on their part in carrying out cognitively complicated tasks. Perhaps some of the perceived deficiencies of survey data trace to exaggerated expectations of this kind. Secondly, however narrowly the research problem may be defined in terms of psychological material, the subject matter is inescapably thoroughly infused by sociocultural definitions and meanings. Quite often, in fact, the main causal variables accounting for the responses recorded in a survey actually are not observed in the survey itself but are to be found in the social milieu, narrowly or broadly construed, of the respondent. [Duncan, 1981:5]

37. In this regard, it is interesting to note that the average user of social survey data knows little or nothing about the interviewers who are the other half of the social interaction that produces these data. While few survey research organizations would fail to provide routine demographic information on respondents, similar information is seldom, if ever, provided about interviewers. Thus, by default, interviewers are treated by most analysts as anonymous and passive transducers of the subjective reality of respondents. It is, we think, a bit odd that as social scientists we often accept such a narrow view of the social realities involved in our own work.

References

Abrams, M. (1973) Subjective social indicators. In United Kingdom (Central Statistical Office), *Social Trends: 1973*. London: Her Majesty's Staionery Office.

Andrews, F., and Withey, S. (1976) *Social Indicators of Well-Being: American's Perceptions of Life Quality*. New York: Plenum Press.

Bailar, B., and Lanphier, C. M. (1977) *Development of Survey Methods to Assess Survey Practices: A Report of the American Statistical Association's Pilot Project on the Assessment of Survey Practices and Data Quality in Surveys of Human Populations*. Washington, D.C.: American Statistical Association.

Blalock, H. (1972) *Social Statistics*. 2nd ed. New York: McGraw-Hill.

Blau, P., and Duncan, O. D. (1967) *The American Occupational Structure*. New York: Wiley.

Boffey, P. (1975) Scientific data: 50 percent unusable; widespread defects in laboratory work found by National Bureau of Standards. *Chronicle of Higher Education* (Feb. 24, 1975):1.

Boring, E. G. (1950) *History of Experimental Psychology*. 2nd ed. New York: Prentice Hall.

Bradburn, N. (1969) *The Structure of Psychological Well Being*. Chicago: Aldine.

Bradburn, N., Sudman, S., and Associates (1979) *Improving Interview Method and Questionnaire Design*. San Francisco: Jossey-Bass.

Bureau of the Census (1980) *Social Indicators III*. Washington D.C.: Government Printing Office.

Campbell, A. (1981) Irregularities in survey data. In D. Johnston, ed., *Measurement of Subjective Phenomena*. Special Demographic Analyses, CDS80-3, U.S. Bureau of the Census. Washington D.C.: Government Printing Office.

Campbell, A., Converse, P., and Rodgers, W. (1976) *The Quality of American Life: Perceptions, Evaluations and Satisfaction*. New York: Russell Sage Foundation.

Caplan, N., and Barton, E. (1976) *Social Indicators 1973: A Study of the Relationship of the Power of Information and Utilization by Federal Executives*. Ann Arbor, Mich.: Institute for Social Research, University of Michigan.

Comptroller General of the U.S.A. (1978) *Better Guidance and Controls are Needed to Improve Federal Surveys of Attitudes and Opinions*. Washington, D.C.: General Accouning Office, Sept. 15.

Cowan, C., Murphy, L., and Weiner, J. (1978) Effects of Supplemental Questions on Victimization Rates from the National Crime Surveys. Paper presented at 138th Annual Meeting of the American Statistical Association, San Diego, Aug. 14-17.

Davis, J. (1975a) Communism, conformity, cohorts and categories: American tolerance in 1954 and 1972-3. *American Journal of Sociology* 81:491-513.

Davis, J. (1975b) Does Economic Growth Improve the Human Lot? Yes, Indeed, About .0005 per Year. Paper presented to International Conference on Subjective Indicators of the Quality of Life, Cambridge, England.

Davis, J. (1976) Background characteristics in the U.S. adult population 1952-1973: a survey-metric model. *Social Science Research* 5:349-383.

Duncan, O. D. (1961) A socioeconmic index for all occupations. In A. Reiss, ed., *Occupations and Social Status*. New York: Free Press.

Duncan, O. D. (1972) Federal statistics, non-federal statisticians. *Proceedings of the American Statistical Association (Social Statistics Section)* 1972:152.

Duncan, O. D. (1974) Developing social indicators. *Proceedings of the National Academy of Sciences* 71:5096-5102.

Duncan, O. D. (1979) Indicators of sex typing. *American Journal of Sociology* 85:251-260.

Duncan, O.D. (1981) Content of surveys. Unpublished ms. prepared for Panel on Survey Measurement of Subjective Phenomena, National Research Council, February 2.

Duncan, O. D., and Schuman, H. (1980) Effects of question wording and context: an experiment with religious indicators. *Journal of the American Statistical Association* 75:269-275.

Easterlin, R. (1974) Does economic growth improve the human lot? Some empirical evidence. In P. Davis and M. Reder, eds., *Nations and Households in Economic Growth*. New York: Academic Press.

Executive Office of the President (Office of Management and Budget). (1973) *Social Indicators: 1973*. Washington, D.C.: Government Printing Office.

Fischhoff, B., Slovic, P., and Lichtenstein, S. (1979) Knowing what you want to know;

measuring labile values. In T. Wallsten, ed., *Cognitive Processes in Choice and Decision Behavior*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Freedman, R., Hermalin, A., and Chang, M. (1975) Do statements about expected family size predict fertility? The case of Taiwan, 1967-1970. *Demography* 12:407-416.

Gibson, C., Shapiro, G., Murphy, L., and Stanko, G. (1978) Interaction of survey questions as it relates to interviewer-respondent bias. *Proceedings of the American Statistical Association (Survey Methods Section)* 1978:251-256.

Goldberg, D., Sharp, H., and Freedman, R. (1959) The stability and reliability of expected family size data. *Millbank Memorial Fund Quarterly* 37:369-385.

Goldfield, E., Turner, A., Cowan, C., and Scott, J. (1977) Privacy and confidentiality as factors in survey response. *Proceedings of the American Statistical Association (Social Statistics Section)* 1977:1-11.

Goodman, L. (1971) The analysis of multidimensional contingency tables: stepwise procedures and direct estimation methods for building models for multiple classifications. *Technometrics* 13:33-61.

Goodman, L. (1972) A general model for the analysis of surveys. *American Journal of Sociology* 77:1035-1086.

Gurin, G., Veroff, J., and Feld, S. (1960) *Americans View their Mental Health*. New York: Basic Books.

Hakel, M. (1968) How often is often? *American Psychologist* 23:533-534.

Hall, J., and Jones, D. (1950) The social grading of occupations. *British Journal of Sociology* 1:31-55.

Hastings, P., and Southwick, J. (1974) *Survey Data for Trend Analysis*. Williamstown, Mass.: Roper Public Opinion Research Center and Social Science Research Council.

Ho, C. Y., Powell, R. W., and Liley, P. E. (1974) Thermal conductivity of the elements: a comprehensive review. *Journal of Physical and Chemical Reference Data* 3(suppl. 1):1-244.

Hunter, J. S. (1977) Quality assessment of measurement methods. In National Research Council, *Environmental Monitoring*, Vol. 4a. Washington, D.C.: National Academy of Sciences.

Kalton, G., Collins, M., and Brook, L. (1978) Experiments in wording opinion questions. *Applied Statistics* 27:149-161.

Kish, L. (1965) *Survey Sampling*. 2nd ed. New York: Wiley.

Kraut, A., Wolfson, I., and Rothenberg, A. (1975) Some effects of position on opinion survey items. *Journal of Applied Psychology* 60:774-776.

Marsh, C. (1978) Opinion polls: social science or political manoeuvre. In J. Evans, J. Irvine, I. Miles, eds., *Demystifying Social Statistics*. London: Pluto Press.

Martin, E. (1978) Trends in victimization: problems of measurement. Paper presented at the 86th annual meeting of American Psychological Association, Toronto, Aug. 28-Sept. 1.

Martin, E. (1981) A twist on the Heisenberg principle: or how crime affects its measurement. *Social Indicators Research* 9:197-223.

Martin, E. (1983) Surveys as social indicators: problems in monitoring trends. In P. Rossi, J. Wright, and A. Anderson, eds., *Handbook of Survey Research*. New York: Academic Press.

Mason, K., Czaijka, J., and Arber, S. (1976) Changes in U.S. women's sex role attitudes. *American Journal of Sociology* 41:573-598.

National Commission on Employment and Unemployment Statistics (1979) *Counting the Labor Force*. Washington, D.C.: Government Printing Office.

National Opinion Research Center (1980) *General Social Surveys, 1972-1980: Cumulative Codebook*. Chicago: National Opinion Research Center.

National Research Council (1979) *Privacy and Confidentiality as Factors in Survey Response*. Washington, D.C.: National Academy of Sciences.

National Science Board (1973) *Science Indicators: 1972*. Washington, D.C.: Government Printing Office.

National Science Board (1975) *Science Indicators: 1974*. Washington, D.C.: Government Printing Office.

National Science Board (1977) *Science Indicators: 1976*. Washington, D.C.: Government Printing Office.

Nisbett, R., and Wilson, T. (1977) Telling more than we can know: verbal reports on mental processes. *Psychological Review* 84:231-259.

OECD (Organisation for Economic Cooperation and Development) (1974) *Subjective Elements of Well-being*. Paris: OECD.

Rich, R. (1975) An Investigation of Information Gathering in Seven Federal Bureaucracies: A Case Study of the Continuous National Survey. Ph.D. dissertation, University of Chicago.

Schuman, H. (1974) Old Wine in New Bottles: Some Sources of Response Error in the Use of Attitude Surveys to Study Social Change. Paper prepared for Research Seminar in Quantitative Social Science, University of Surrey, April 1974.

Schuman, H., and Duncan, O. D. (1974) Questions about attitude survey questions. *Sociological Methodology: 1973-74*. San Francisco: Jossey-Bass.

Schuman, H., and Johnson, M. (1976) Attitudes and behavior. In A. Inkeles, ed., *Annual Review of Sociology* 2:161-207.

Schuman, H., and Presser, S. (1977) Question wording as an independent variable in survey analysis. *Sociological Methods and Research* 6:151-170.

Schuman, H., and Presser, S. (1978) The Assessment of "No Opinion" in Attitude Surveys. Paper presented at 73rd annual meeting, American Sociological Association, San Francisco, Sept. 4-8.

Sewell, W., and Hauser, R. (1975) *Education, Occupation, and Earnings*. New York: Academic Press.

Sheldon, E. (1971) Social reporting for the 1970's. In Presidential Commission on Federal Statistics, *Federal Statistics*, Vol. 2. Washington, D.C.: Government Printing Office.

Shure, G., and Meeker, R. (1978) A minicomputer system for multiperson computer-assisted interviewing. *Behavior Research Methods and Instrumentation* 10:196-202.

Smith, T.W. (1978) In search of house effects: a comparison of responses to various questions by different survey organizations. *Public Opinion Quarterly* 42:443-463.

Smith, T.W. (1979a) *Can We Have Confidence in Confidence? Revisited*. GSS Technical Report No. 10. Chicago: National Opinion Research Center.

Smith, T.W. (1979b) Happiness: time trends, seasonal variations, intersurvey differences, and other mysteries. *Social Psychology Quarterly* 42:18-30.

Staines, G., and Quinn, R. (1979) American workers evaluate the quality of their jobs. *Monthly Labor Review* 102:3-12.

Steiner, E. (1975) Planning and analysis of the results of collaborative tests. In W. Youden

and E. Steiner, eds., *Statistical Manual of the Association of Official Analytical Chemists*. Washington, D.C.: Association of Official Analytical Chemists.

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

Treiman, D. (1977) *Occupational Prestige in Comparative Perspective*. New York: Academic Press.

Turner, C. F. (1981a) Surveys of subjective phenomena: a working paper. In D. Johnston, ed., *Measurement of Subjective Phenomena*. Special Demographic Analyses, CDS80-3, U.S. Bureau of the Census. Washington D.C.: Government Printing Office.

Turner, C. F. (1981b) Patterns of disagreement: a reply to Angus Campbell. In D. Johnston, ed., *Measurement of Subjective Phenomena*. Special Demographic Analyses, CDS80-3, U.S. Bureau of the Census. Washington D.C.: Government Printing Office.

Turner, C., and Krauss, E. (1978) Fallible indicators of the subjective state of the nation. *American Psychologist* 33:456-470.

United Kingdom (Central Statistical Office) (1970-) *Social Trends*. London: Her Majesty's Stationery Office (annual).

United Nations (Department of Economic and Social Affairs) (1975) *Toward a System of Social and Demographic Statistics*. New York: United Nations.

U.S. Department of Commerce (1977) *Social Indicators: 1976*. Office of Federal Statistical Policy and Standards. Washington, D.C.: Government Printing Office.

Waksberg, J. (1975) How good are survey statistics? *Proceedings of the American Statistical Association (Social Statistics Section)* 1975:26-27.

Wilson, F., and Bumpass, L. (1973) The prediction of fertility among Catholics: a longitudinal analysis. *Demography* 10:591-597.

Wilson, T., and Nisbett, R. (1978) The accuracy of verbal reports about the effects of stimuli on evaluations and behavior. *Social Psychology* 41:118-130.

Youden, W. (1975) Statistical techniques for collaborative tests. In W. Youden and E. Steiner, eds., *Statistical Manual of the Association of Official Analytical Chemists*. Washington, D.C.: Association of Official Analytical Chemists.

Acknowledgments

The author is grateful for the helpful comments provided by a number of people including Robert P. Abelson, Donald Buzzelli, Clifford Clogg, Otis Dudley Duncan, Sara Kiesler, Denis Johnston, Cathie Marsh, Elizabeth Martin, Naomi D. Rothwell, Howard Schuman, and Tom Smith. An early draft of this chapter was published as a working paper (Turner, 1981a); an exchange of views with Angus Campbell (1981; Turner, 1981b) doubtlessly strengthened this final product.