FROM: D. Johnston (Editor), Measurement of Subjective Phenomena.
Washington D.C.: U. S. Government Printing Office, 1981.

Surveys of Subjective Phenomena: A Working Paper

Charles F. Turner
National Research Council
National Academy of Sciences

OVERVIEW

In a broad sense, the present chapter is concerned with the interrelation of psychology and survey research. Our more particular concerns are prompted by anomalies in the results of surveys that purport to measure identical subjective phenomena in a comparable manner. The burden of these anomalies, we submit, is sufficiently weighty to motivate a reconsideration of the psychological assumptions underlying the practice of survey research.

In the present chapter we assemble a variety of new evidence and suggest some tentative hypotheses concerning the types of subjective phenomena particularly vulnerable to artifacts of measurement. The present article is a montage; it presents abbreviated analyses of a number of examples rather than intensive case studies. The future, we hope, will provide the time and resources for a careful reconsideration of each of these examples.

AREA OF INQUIRY

Traditionally the term *subjective* has been used to denote those phenomena which 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 that although frequently measured by subjective means (i.e., self-report) are theoretically amenable to independent confirmation. 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 (i.e., objective) reality whose

This working paper is based on a talk delivered at the 86th annual convention of the American Psychological Association in Toronto, Canada, August 28-September 1, 1978. The views expressed herein are the sole responsibility of the author; they should not be attributed to the National Academy of Sciences or the National Research Council.

properties are potentially discoverable through some consensual process. (One, of course, need not make such an assumption; see, e.g., the writings of Bishop Berkeley or the radical skepticism of Descartes' first meditation.)

So, 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 (e.g., educational attainment, geographic mobility, fertility history, family structure, income). However, many important phenomena are inherently subjective and thus immune to third-party verification. In particular, we have no direct knowledge[1] of an individual's attitudes, beliefs, or opinions. The present inquiry focuses upon survey measurements of such *subjective* phenomena.

IMPORTANCE OF AREA

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 third-party[2] corroboration, for example, age, income, educational attainment. 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 supported by studies using independent sources (e.g., birth and earnings records) to estimate the magnitude and sources of error introduced by the exclusive use of information supplied by subjects themselves (see, e.g., the work of the evaluation and research program of the Census Bureau).

Our concern centers upon several 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 statistics, contain fundamentally subjective elements. For example, to be classi-

---

[1] Observation of a subject's behavior can only provide an indirect measure of a subject's attitudes, beliefs, etc. Conceptually, one can distinguish between behaviors and attitudes; in practice, experience affirms that an individual's behavior need not accord with beliefs. Indeed, one should bear in mind that typical measurements of subjective states involve inferences made from a subject's behavior(s), i.e., 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, e.g., the recent review by Schuman and Johnson, 1976).

[2] In this regard, we note the traditionally unproblematic concept of a *true population value* in studies of sampling and response error in census work. This concept becomes more difficult to sustain (at least intuitively) when discussing the measurement of subjective phenomena (cf. Waksberg, 1975).

fied as unemployed in the U.S. Bureau of Labor Statistics (BLS) index, workers must be both out of work and looking for work. While the definition and interpretation of the criterion, looking for work, are not left solely to the discretion of respondents, statements that a worker has "checked with friends or relatives" during the last 4 weeks satisfy the second criterion. This component of the unemployment index is fundamentally subjective because the meaning and interpretation of the statement, "checked with friends . . . ," is supplied by the respondents and is inherently unverifiable. Similarly, evaluations of the National Crime Surveys by the U.S. Bureau of the Census (cf. Cowan et al., 1978; Gibson et al., 1978) have pointed up the subjective components of crime victimization statistics derived from subjects' self-reports. In the latter instance, the meaning and interpretation of phrases 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 the measurements are made (cf. Martin, 1978, 1981).

In recent years, national statistics have come to include an important and rapidly growing complement of statistics designed to measure subjective phenomena. For example, the *Social Indicators* (U.S. Department of Commerce, 1973, 1977) program begun by the Office of Management and Budget incorporates measurements of a wide range of subejctive phenomena. (An independent review (Caplan and Barton, 1976) of the use of the first *Social Indicators* volume concluded that there was a need for Federal statistical compendia to "go beyond objective indicators and provide subjective measures of life experience and social well-being.") The most 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 members. The bulk of the information presented relates to people's objective situation or condition—their jobs, their income, 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 on people's values and aspirations. (U.S. Department of Commerce, 1977, p. 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 fact that financial support, the imposition of legal constraints (e.g., regulation of recombinant DNA research), and the recruitment of young people into the scientific professions depends, in part, upon public perceptions of science. In the Board's own words,

> 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. (National Science Board, 1975, p. 145)

The increasing importance of measures of subjective phenomena in Federal statistical programs is paralleled by a growing range of relevant research activities[3] in the academic community. This work has included psychological studies of well-being (e.g., Campbell et al., 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 (e.g., Davis, 1975a; Duncan, 1979; Mason et al., 1976), and work by economists on the relationship of economic development to individual happiness (e.g., Easterlin, 1974).

Similar trends are evident around the world. Since 1970, Britain has issued annual reports entitled *Social Trends.* These reports pay considerable attention to measures of subjective phenomena. Indeed, they argue that:

> The more one considers [indicators of the quality of life], the more one is persuaded that the way forward lies not in adding more measures of conventional hard statistics, but rather in supplementing the existing ones by adding . . . a dimension of the satisfaction (happiness, contentment, psychological well-being, etc.) felt by those who constitute the community and are the final consumers of society's "goods" and "bads" and therefore the best judges of society's performance. In short it is the very thoroughness [of earlier work on hard statistics] that compels one to turn to subjective social indicators and to the problems of reliable quantification of states of mind and mood. . . . (Abrams, 1973, p. 36)

Moreover, a recent United Nations report, *Toward a System of Social and Demographic Statistics* (1975), echoes this concern for the inclusion of measures of subjective phenomena in national statistics. In a chapter beginning,

---

[3]The emergence of such efforts is reflected by a number of occurrences. For example, the social science community has established a national data program with the specific aim of providing data from representative national samples to enable the construction of long term time series of both demographic and attitudinal indicators. Similarly, there was a unique, albeit short-lived, attempt to provide Federal policymakers with easy access to an ongoing series of national surveys. The latter project was funded by the RANN division of the National Science Foundation (NSF) for the use of the following agencies: Department of Agriculture; Department of Health, Education, and Welfare; Department of Housing and Urban Development; Department of Transportation; Office of Management and Budget; President's Commission on Gambling; and NSF. For a history of this pioneering but ill-fated program, see Rich (1975).

"The need for social indicators . . .," the report observes that

> In dealing with social questions, however, we may also be interested in subjective information relating to how much people in general know about an issue, how much importance they attach to it and what kind of solution they think would be desirable.
>
> Public opinion surveys provide a means of obtaining some light on such matters. It would be interesting to know, for instance, what issues are commonly regarded as major problems and how the ranking of these issues changes with time. It would also be interesting to know how far the public connects one issue with another: does it believe, rightly or wrongly, that the great increase in pollutants in recent years is associated with activities and processes that contribute to the rising standard of living; does it believe, rightly or wrongly, that the scale and organization of modern enterprise, which also contribute to the standard of living, are associated with industrial unrest and alienation?

COMPARABILITY OF SUBJECTIVE SOCIAL INDICATORS

Survey measurements of subjective phenomena are made by many organizations. In the United States, nonfederal sources produced the majority of the subjective social indicator measurements reported in *Social Indicators; 1976.*

The use of nongovernmental surveys to collect such attitudinal information reflects both ideological and practical considerations. It is sometimes argued that on principle governments ought not to ask their citizens for certain information even though it might be useful in developing government policy. Concerning religious beliefs, for example, it has been argued that "no federal statistical agency has any business whatever inquiring about anyone's religious beliefs, even though information about the distribution of beliefs in the population is pertinent to various federal policies" (Duncan, 1972, p. 152).

At a practical level, doubts have been raised about the ability of governments to obtain reliable information on sensitive topics, for example, antigovernment sentiments.[4] Thus, in an appendix to the report of the Presidential Commission on Federal Statistics, Sheldon (1971) proposed a division of labor between government and nongovernment statistical organizations,

> The development of time series information covering subjective dimensions as well as topics presumed to be politically sensitive will continue to

---

[4] A recent survey conducted under the auspices of the Committee on National Statistics (National Academy of Sciences) suggests that such concerns may be overstated. In a survey conducted jointly by the Census Bureau and the Survey Research Center of the University of Michigan, virtually identical results were obtained by both organizations on a large variety of items, including several measuring antigovernment attitudes. The only significant differences occurred for items specifically asking about social surveys and survey-taking organizations (see Goldfield et al., 1977; National Academy of Sciences, 1979).

> be the primary responsibility of non-governmental research and university centers. . . . This work requires considerable conceptual innovation and field experimentation activities particularly appropriate to institutions independent of governmental agencies. (p. 421)

Use of data from a variety of sources, however, inevitably raises questions of comparability. Despite one's hopes, comparability of measurement does not occur naturally (cf. Hunter, 1977; Ho et al., 1974); experience indicates that it is, rather, the result of careful standardization of research procedures and the continuous monitoring of performance. For example, attempts by analytical chemists to achieve comparability of measurement across laboratories involved a long history of standardization of research practices and the development of methods for collaborative tests *across laboratories* (cf. Youden and Steiner, 1975).

We believe that emerging evidence of noncomparability in survey measurements of subjective phenomena argues both for a consideration of appropriate techniques for assuring data comparability when measurements are derived from several sources and a reconsideration of the psychological assumptions that underlie the practice of survey research.

THE PROBLEM

The use of replicated time series of subjective social indicators has given birth to some disagreeable progeny. Most irritating has been the multiplication of instances in which allegedly identical measurements of subjective phenomena have differed both substantially and significantly between surveys (cf. Turner and Krauss, 1978; Martin, 1981); in some instances, discrepancies of 15 percentage points in the univariate distributions have been observed. These discrepancies prompt a number of questions; one would like to know for example:

Why these measurements disagree
Whether these disagreements are symptomatic of a larger problem or whether they are restricted to a few isolated cases
Whether there are any organizing principles that could provide a typology of those indicators more (and less) likely to give discrepant results

In the following pages we review examples drawn from a variety of social indicator projects. These examples shed some light upon these issues and illustrate the need for further research. In addition, we test some initial organizing principles concerning the types of subjective indicators that are particularly vulnerable to artifacts of measurement. We do so not in the hope of elucidating final principles but rather to provide initial hypotheses around which research may be organized. As a starting point, we take the recent suggestion (Turner and Krauss, 1978, p. 468) that discrepancies in estimates of subjective social indicators may be a function of the survey questions, them-

selves, and the phenomena they intend to measure. In particular, it was hypothesized that the discrepancies would be concentrated among those indicators that involved survey questions that

1. Were most amorphous in their meaning, for example, those seeking to assess "confidence," "trust"
2. Were most ambiguous in their referents, for example, those inquiring about the "people running organized religion"
3. Involved the most arbitrariness in the selection of a response category, for example, great deal versus some confidence

Using these hypotheses as our point of departure, let us turn to our first example.

THE CASE OF HAPPINESS

This example is the first of several where surveys disagree. 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 sensitive estimates of life satisfaction (e.g., Gurin et al., 1960; Bradburn, 1969; Campbell et al., 1976). Much 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 (cf. Gurin et al., 1960), and both the trends across time and differences between nations in response to this question have been discussed by several authors (e.g., Easterlin, 1974; Davis, 1975b; Campbell et al., 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 1 presents two independent series of happiness estimates derived from surveys conducted by the Survey Research Center of the University of Michigan (SRC) and the National Opinion Research Center of the University of Chicago (NORC). This figure shows that there are not only discrepancies in estimates of the absolute *levels* of happiness but also that the trends in the two series appear to diverge in direction. One series shows an apparent increase while the other series registers a decline in happiness.

We first noted 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

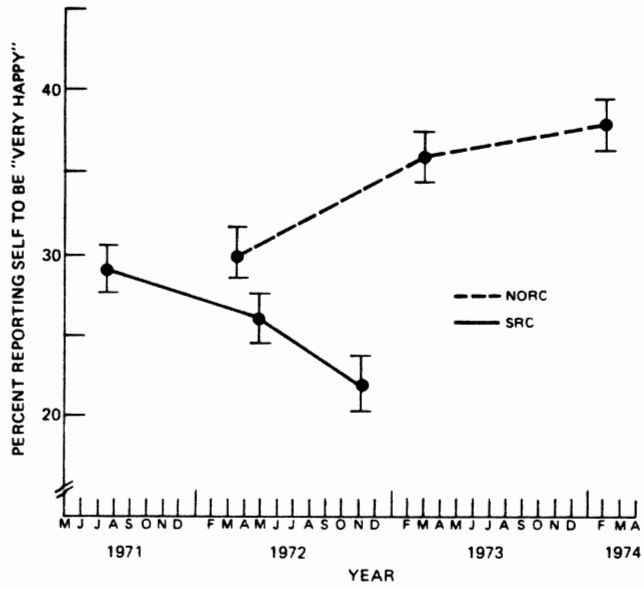


Figure 1. Trends in self-reported happiness, 1971-1973. Estimates are derived from sample surveys of noninstitutionalized population of the continental United States, aged 18 and over. Error bars demark ±1 standard error around sample estimate. *Source:* NORC, *National Data Program for the Social Sciences: Codebook, 1972-1974.* SRC, estimates from Campbell et al. (1976); survey dates from Campbell et al. (1976), Andrews and Withey (1976), and J. Varva (personal communication).

the discrepancies that had been observed in the "confidence in institutions" series.[5] Subsequent examination of the two happiness series has caused us to doubt the validity of the comparison shown in figure 1. Examination of the questionnaires used by NORC and SRC reveals some differences in question wording:

> 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)

The SRC version of the happiness question repeats "these days" at the end of the question, whereas the NORC version does not. Thus, it might be argued that these questions were indicators of slightly different phenomena, although

[5]The meeting was hosted by the Institute for Research in the Social Sciences, University of North Carolina, and attended by J. Davis, O. D. Duncan, E. Martin, F. Munger, R. Parke, 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 (cf. T. Smith, 1979).

admittedly one might expect the trends across time to be parallel rather than divergent. However, it could be that the divergences are more apparent than real. For example, national happiness might have fluctuated rapidly between 1971 and 1973, and thus, these data could be reliably mirroring *month-to-month* changes 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 (e.g., Campbell et al., 1976, p. 26; Andrews and Withey, 1976). Moreover, large month-to-month fluctuations in these indicators would preclude use of annual and biennial reporting schedules, which have been the rule in the social indicators field. (If the trends are month-to-month, then yearly or biennial measurements (taken in different months) may prove uninformative or misleading.)

Thus, despite their limitations, the data shown in figure 1 prompted us to speculate about the 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.

We hypothesized that insertion of this marital happiness question created an artifactual response bias. Our initial examination of this hypothesis (tables 1 and 2) indicated that:

1. There was a high correlation between responses to the marital and general happiness questions
2. The marital happiness question elicited a relatively high proportion (.6) of very happy responses
3. The increase in overall happiness between 1972 and 1973-1974 in the NORC series occurred only among married persons.

This last finding[6] was particularly important because the hypothesized context effect could only have occurred for married individuals. Unmarried persons were not asked, of course, about the happiness of their marriages.

While any comparison of the NORC and SRC happiness series admits to a plethora of alternate explanations (e.g., wording effects, house effects, short term temporal variations), the results of our initial explorations encouraged us

[6] 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 X marital status) is not strictly required.

Table 1. Association Between Responses to Marital Happiness and General Happiness Questions: 1973-1977

Marital happiness	General happiness (percent)			N^a
	Not too happy	Pretty happy	Very happy	
Not too happy	65	32	3	150
Pretty happy	11	78	11	1,502
Very happy	5	38	57	3,408

Note: $\chi^2 = 1{,}094.6$; df = 4; $p < .0001$; $\gamma = +0.75$

[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 1973-1978 $r_i = 0.02$) for the happiness item indicates that this correction provides a conservative estimate of the relevant sampling errors. Sample sizes shown are raw figures; they do not reflect deflation or weighting of sample by number of eligible adults in household.

to seek a better test for our hypothesis. We were furtunate to discover a wealth of information on happiness collected during this period. Between April 1973 and May 1974, the National Opinion Research Center with the support of the RANN division of the National Science Foundation conducted a series of pilot surveys to provide continuous monitoring of public opinion for policymakers in eight Federal agencies. At intervals of approximately 1 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 (CNS).

Table 2. Married and Unmarried Respondents Reporting Themselves To Be Very Happy: 1972-1974

Sample	1972 (percent)	1973 (percent)	1974 (percent)	χ^2 for temporal change[a]	
				1972 vs. 1973 vs. 1974[b]	1972 vs. 1973 + 1974[c]
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 ($p < .001$)

NS Not significant.

[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 1973-1978 $r_i = 0.02$) for the happiness item indicates that this correction provides a conservative estimate of the relevant sampling errors. Sample sizes shown are raw figures; they do not reflect deflation or weighting of sample by number of eligible adults in household.

[b]df = 2.

[c]df = 1.

These data allow us to compare responses across time for two identically worded questions in surveys conducted by the same research organization. This provides a control for both wording differences and any possible organizational idiosyncracies (e.g., 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.

Figure 2 presents a graphic summary of our findings. Specifically, we found that for unmarried individuals, yearly estimates derived from the NORC General Social Survey (GSS) and the monthly estimates from the CNS were in general agreement. This is not to say that the estimates were identical. However, observed discrepancies were well 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 1972-1974 GSS and the 12 cycles of the CNS.

For the married respondents, a rather different result emerged. In particular, the GSS happiness estimates exhibit a sharp rise ($\chi^2 = 21.5$; df = 1; $p < .0001$) between 1972 and 1973-1974 (change = +10 percent), but the 12 monthly

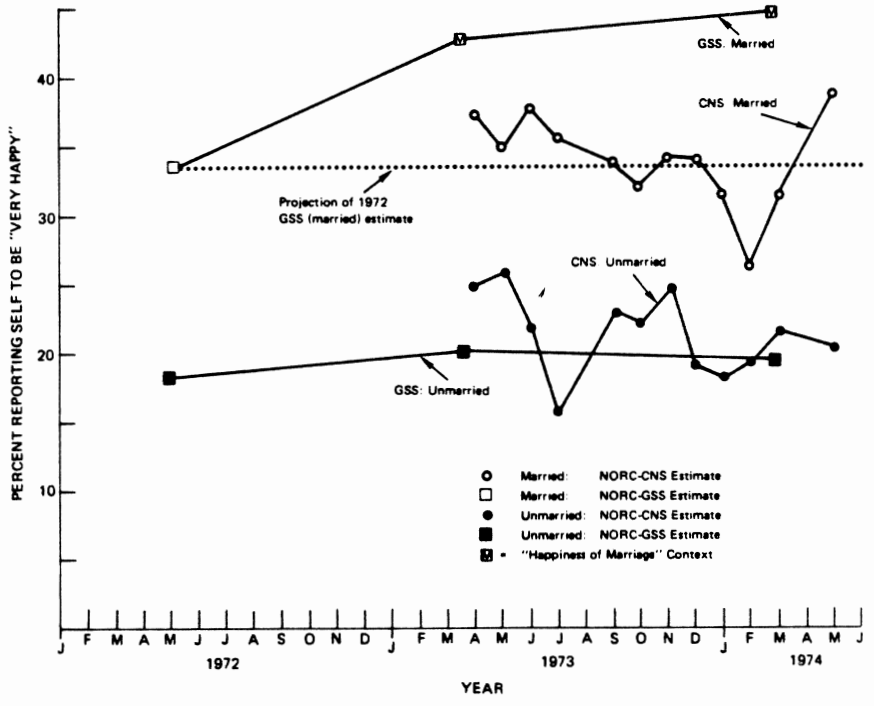


Figure 2. Variations in response to NORC happiness question for married and unmarried respondents in the General Social Surveys (GSS) and Continuous National Surveys (CNS). Estimates are derived from samples of approximately 1,000 (GSS) and 440 (CNS) married respondents and 500 (GSS) and 220 (CNS) unmarried respondents.

estimates derived from the CNS evidence no similar trend. Moreover, the CNS happiness estimates are consitently below those of the GSS (average difference = 9 percent). Indeed, as figure 2 shows, the 1972 GSS measurement—when the marriage question was not included—provided a better prediction of the CNS estimates in 1973 and 1974 than the actual GSS estimates in those years.

Although other hypotheses might be supported, we submit that (1) the internal evidence of a temporal trend *only* for married GSS respondents and, (2) the predictability of the 1973-1974 CNS data from 1972 GSS data provide strong support for the hypothesis that a response bias arose from the insertion of the marital happiness question in the 1973 and 1974 General Social Surveys.

The implications of this artifact are substantial. The GSS provides a major source of data for the social science community, and it was the largest source of subjective social indicator data for the Federal compendium, *Social Indicators, 1976* (U.S. Department of Commerce, 1977). It is not unreasonable to anticipate publication of substantive interpretations for the post-1972 rise in GSS happiness series (e.g., as an effect of the end of the Vietnam war). Such interpretations, of course, would be misleading. This increase in national happiness appears to result from changes in the content of our surveys and not from changes in the subjective state of the national population.

What do we learn from this example? Recalling the general hypotheses outlined in the introduction, we observe that the

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

FERTILITY EXPECTATIONS

Lest the reader be misled by our first example, we hasten to note that we do not believe that all survey measurements of subjective phenomena are equally vulnerable to artifactual biases. Rather, we wish to delineate areas in which artifact-induced discrepancies might be expected and the factors likely to cause such misbehavior. With this purpose in mind, let us consider some alternative estimates of the fertility expectations of American women.

The U.S. Bureau of the Census conducts annual surveys of the birth expectations of American women. The data from such surveys are potentially useful in predicting fluctuations in the birth rate. (Longitudinal studies indicate that those measures do predict subsequent fertility; e.g., Wilson and Bumpass, 1973; Freedman et al., 1975; Goldberg et al., 1959.) Clearly, the phenomenon being measured in such surveys is highly subjective. The intention or expectation of

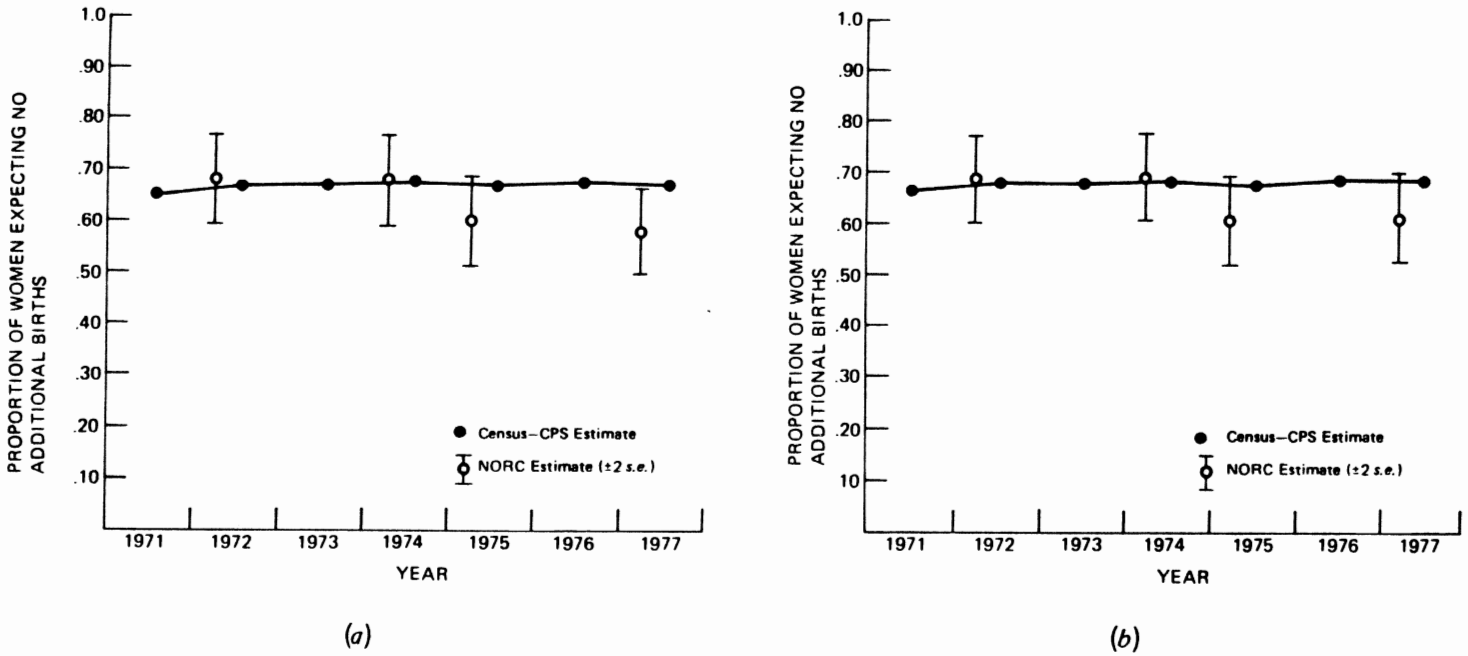


Figure 3. Estimates of fertility expectations of American women: proportion of women expecting no further children in (*a*) all future years and (*b*) next 5 years. Samples included only married women aged 18 to 39; sample sizes in each year were approximately 4,000 (Census-CPS) and 220 (NORC).

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 Census Bureau's Current Population Survey (CPS) are shown in figure 3 together with estimates derived from a related question on birth expectations asked in the 1972 and 1975-1977 NORC GSS. It should be noted that the latter estimates are based on a very small sample; on the average, there were fewer than 250 married women aged 18-39 in the GSS samples. Thus, the standard errors for the GSS estimates are quite large (about 4 percent).

Comparing the two sets of data, we find that estimates of fertility expectations derived from NORC's GSS are quite consistent with those derived from the Census Bureau's CPS. In only one instance (of eight) does the NORC GSS estimate differ by more than two standard errors from the Census estimate.

The Lesson

The estimates of fertility expectations presented in figure 3*a* and 3*b* 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. Table 3 presents the actual text of the questions used in each measurement. The questions asked about birth expectations varied slightly not only between the NORC and Census series but also within the Census series across time (i.e., 1972 vs. 1973-1977).

Because the comparison of these two series of birth expectations involved both wording and context differences and because 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?

Table 3. Birth Expectations Questions Used by Census Bureau and NORC: 1972-1977

Organization	Years	Text
Census Bureau	1972-1974	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 five years?
	1975-1977	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 five years?
NORC	1972-1977	1. Do you expect to have any (more) children? 2. How many (more)? 3. How many (more) in the next five years?

In terms of our initial hypotheses, we note:

1. The birth expectation question is relatively unambiguous in its meaning.
2. The response categories for the questions (e.g., 0, 1, 2, 3, . . . children) have a rather clear meaning.
3. The question deals with a topic to which most respondents (i.e., married women of childbearing age) should have given considerable thought, particularly because attitudes toward childbearing have behavioral consequences in the everyday life of the respondents, for example, contraceptive behaviors.

SCIENCE AND THE PUBLIC

Our next two examples of disagreement 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) of Princeton, N.J. 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.

Our interest in these surveys was first aroused by an observation made during the analysis of the 1976 survey. In brief, the 1976 survey contained an anomaly that particularly concerned the staff person responsible for the chapter on public attitudes toward science. This anomaly had potentially destructive implications for national science policy and funding; thus, it is a most appropriate illustration of the dangers inherent in our inadequate understanding of the error structure of the data we employ as subjective social indicators.

The anomaly in the 1976 survey arose through an attempt to explore the meaning of public response to the following questions:

> 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 gave relatively strong endorsement to funding science in order to reduce crime (59 percent in 1972 and 58 percent in 1974), fight drug addiction (51 percent and 48 percent) and 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 question in the 1976 survey.

In 1976, the Science Indicators survey was altered to incorporate the following questions *immediately prior* to the spending questions:

> 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 above)
>
> In your view, in which areas could science and technology make little or no contribution? Please read me the numbers.

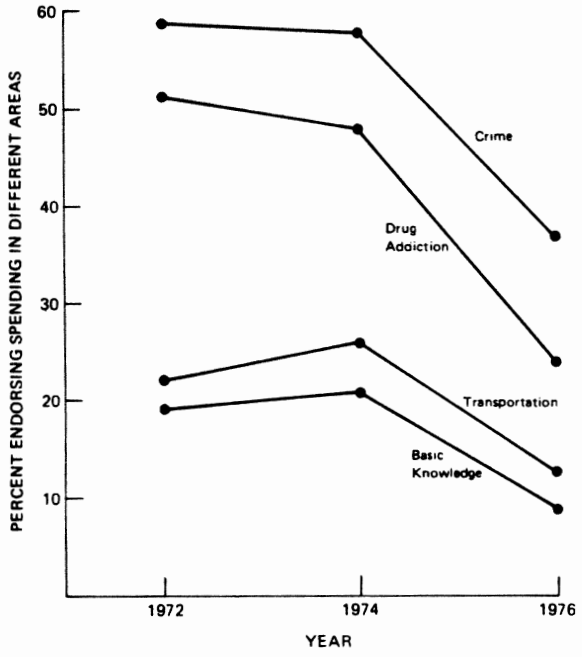


Figure 4. Endorsement of spending for science and technology in four areas. Sample size in each year was approximately 2,100. *Source:* Science Indicators surveys, 1972-1976.

Table 4. Endorsement of Spending for Science and Technology, Estimated by Science Indicator Surveys: 1972, 1974, and 1976

Area	Endorse spending (percent)			Change	
	1972	1974	1976	1972-1976	1974-1976
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	34	34	21	−13	−13

Note: Estimates are for percent selecting areas in which they would most like to have taxes spent. See text for question wording.

Responses to the spending question in the 1976 survey were so unusual, however, that neither an analysis of the relationship between the perceived usefulness of science and the endorsement of spending, nor the spending time series itself appear in the final report of the National Science Board. Instead, a footnote observed that alterations in the ordering and content of the questions preceding the spending question preclude a valid comparison of the 1976 estimates to those obtained in previous years. This reticence is understandable.

Figure 4 and table 4 present the estimates derived from responses to the spending question in 1972, 1974, and 1976. These estimates show an apparently precipitous decline in public support of spending for science and technology. In two instances, this decline exceeded 20 percentage points.

This evidence of a massive drop in public support is, however, inconsistent with other independent evidence. The GSS has included an item on spending since 1973. The resultant NORC series show virtually constant levels of public support for science-related spending between 1973 and 1976 (table 5).

Because of the change in the questionnaire used in the Science Indicators surveys, the National Science Board chose not to present the data showing an apparent decline in public support for science spending. Instead, they argued in a footnote that

Table 5. Evaluation of Government Spending Programs, Estimated by NORC General Social Surveys: 1973-1976

Area	Spending too little or about right (percent)			Change	
	1973	1974	1976	1973-1976	1974-1976
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: Estimates are repercentaged to exclude "don't know" responses and no answers. Sample sizes in each year were approximately 1,500.

Question: 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 _____ ?

> 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, p. 180)

Our own analysis of the NORC spending data indirectly supports this argument. However, this position contradicts the prevailing wisdom among survey researchers. In their well-known book, *Response Effects in Surveys*, Sudman and Bradburn (1974) concluded 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. (p. 33)

Context and Univariate Distributions

The National Science Board argued that their 1976 estimates of public endorsement of science spending were not comparable to those in previous years because of the different questionnaire contexts in which the spending question was embedded. Sudman and Bradburn argue that, in general, the effects of such context variations are negligible. Who is right?

In the absence of experimental evidence upon the specific context effects postulated by the National Science Board, it is difficult to assess the validity of their claim. The stability of public support for spending in other independent series circumstantially supports the Board's position. Similarly, recent experimental evidence (cf. Turner and Krauss, 1978, p. 466 ff.) indicates that context variations may produce fluctuations of up to 15 percentage points in the univariate distributions of some measurements.

While we cannot directly test the claim made by the Board, we have studied other circumstantial evidence on this question. 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 1976 report were asked along with questions on hospitalization and medical expenditures, frequency of eating hamburgers, and the litter problem.

To control for context effects in these surveys, two different versions of the survey were administered. The different versions rotated the order of questionnaire sections in the questionnaire. Also, for some multipart questions, the sequence of individual parts of a question was varied. Each version of the questionnaire was administered to one-half of the sample.[7]

The first item in the Science Indicators section of the 1976 questionnaire assessed the prestige, or general standing, of various occupations including scientist and engineer. For the National Science Board, this question is of interest both because it serves as a surrogate measure of public attitudes toward

[7] Selection of respondents to receive form A or B took place as follows: The survey, itself, used three separately generated national samples. In sample 1, 60 sampling points received version A and 60 received B; selection was random. 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.

Randomization by sampling points (Ns = 8) 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 in this instance is impossible to compute given the presently available data. In the analysis of the confidence items, Turner and Krauss (1978) 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) (cf. H. Blalock, 1972, chapter 20; Kish, 1965).

Whether that result holds here is unknown; however, we would point out that anlaysis of the demographic characteristics of the samples (age, sex, education, income, number in household, religion, marital status) by form (A or B) revealed no significant differences (computations assume effective sample size to be deflated by 0.66 (i.e., N = 2,000 becomes N = 1,333)

Variable	χ^2	df	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

science and because the prestige of scientific occupations influences recruitment of talented young people into these professions. For social scientists, the responses to such questions are important because they provide the basis for well-known scalings of the socioeconomic status or prestige of occupations (Duncan, 1961; Hall and Jones, 1950; Treiman, 1977). These scales have been central to much recent work on social stratification (e.g., Blau and Duncan, 1967; Sewell and Hauser, 1975).

It is thus of considerable interest to know whether response to this question was affected by the context variation built into the Science Indicators survey. The survey question 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 shown a card containing the responses: excellent, good, average, below average, and poor. Ratings were solicited for the 10 occupations shown in table 6.

The variation in the context and administration of the prestige question was twofold. First, the order of occupations listed in form A was reversed in form B (i.e., from businessman through accountant in form A, and in the reverse order in form B). Second, the placement of this question in the survey varied. In

Table 6. Variations in Excellent Ratings of Occupational Prestige, by Survey Form

Occupation	Excellent (percent)		Discrepancy (percent)	x^2	p^a
	Form A	Form B			
Businessman	13.4	13.4	0.0	0.0	NS
Physician	47.6	56.4	−8.8	17.8	.005
Scientist	46.8	49.5	−2.7	8.4	NS
U.S. Representative in Congress	16.1	30.4	−14.3	54.1	.0001
Lawyer	24.0	38.7	−14.7	41.9	.0001
Architect	24.6	37.5	−12.9	29.9	.0001
Minister	39.0	38.3	+0.7	9.3	NS
Engineer	25.5	34.0	−8.5	18.9	.002
Banker	18.9	27.7	−8.8	26.7	.0001
Accountant for a large business	17.3	25.0	−7.7	28.9	.0001

Note: Listing of occupations is in order used in form A; the reverse order was used in form B. Wording of occupational titles is identical to that in questionnaire. Chi-square tests were performed across the entire response distribution (i.e., excellent, good, average, below average, poor, and no opinion); the degrees of freedom for the tests were 5. To conserve space, only the distributions for the excellent response category are shown; this category accounted for a majority of the variability across forms.

NS Not significant.

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

form A, this question was the very first question in the survey. The interviewer was instructed to begin with a 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 proceeded to the special introduction required for 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 proceeded to ask a series of 38 questions on litter, hamburger makers, and hospitalization and medical insurance.[8] Following these questions, the interviewer delivered the NSF introduction and proceeded to ask the question on the prestige of occupations.

The divergence in the results obtained from these two different administrations is striking. Table 6 presents the relevant comparisons. For 8 of 10 occupations, rated prestige is lower when the question is asked at the beginning of the survey (form A). For 7 of the 10 occupations, this difference is 5 or more percentage points, and in four cases, it exceeds 10 percentage points. The sole

---

[8] This survey, the ORC Caravan, consisted of eight parts, each funded by different organizations. The Opinion Research Center 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 lists derived from summaries prepared by ORC staff describe the content of survey sections.

1. Seriousness of litter problem in U.S.
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

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.

1. A series of questions on hospitalization insurance, family members covered, etc.
2. Proportion of expenses paid by such coverages
3. Hospitalization incidence last year
4. Estimated cost of hospitalization

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 on the two forms of the survey. Respondents who were first exposed to questions on litter, hamburger makers, and hospitalization gave generally more favorable evaluations of these 10 professional occupations. Why this happened is unclear. One might speculate that survey respondents have an initial set against the use of extreme response categories (e.g., excellent). This bias may diminish with practice in responding to survey questions. However, some experimental evidence suggests a modest trend in the opposite direction (cf. Kraut et al., 1975). Alternatively, one might speculate that sequencing banal questions on beverage containers, litter, and hamburger makers, immediately before questions about acute medical problems

Table 7. 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		Discrepancy	χ^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

Note: This analysis focuses attention upon the "great deal of confidence" category in accord with common reporting practices (cf. *The Harris Survey*, December 6, 1973; September 30, 1974; October 6, 1975; March 22, 1976; March 14, 1977; January 5, 1978). We have eliminated missing data ("don't know," no answer, etc.) from the response distributions for each item. Chi-square statistics have one degree of freedom and are corrected for continuity (Yates correction). Given that assignment to experimental conditions was fully random, the analysis treats the respondents (average sample size about 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?

NS Not significant.

[a]In order of presentation to respondents.

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 on the two forms of the survey. Respondents who were first exposed to questions on litter, hamburger makers, and hospitalization gave generally more favorable evaluations of these 10 professional occupations. Why this happened is unclear. One might speculate that survey respondents have an initial set against the use of extreme response categories (e.g., excellent). This bias may diminish with practice in responding to survey questions. However, some experimental evidence suggests a modest trend in the opposite direction (cf. Kraut et al., 1975). Alternatively, one might speculate that sequencing banal questions on beverage containers, litter, and hamburger makers, immediately before questions about acute medical problems

Table 7. 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		Discrepancy	χ^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

Note: This analysis focuses attention upon the "great deal of confidence" category in accord with common reporting practices (cf. *The Harris Survey*, December 6, 1973; September 30, 1974; October 6, 1975; March 22, 1976; March 14, 1977; January 5, 1978). We have eliminated missing data ("don't know," no answer, etc.) from the response distributions for each item. Chi-square statistics have one degree of freedom and are corrected for continuity (Yates correction). Given that assignment to experimental conditions was fully random, the analysis treats the respondents (average sample size about 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?

NS Not significant.

[a]In order of presentation to respondents.

and experience with doctors and hospitals, created a proscience and proprofessional evaluation bias. The latter speculation may be plausible, especially because respondents were told that they were being asked to evaluate these professional occupations for "the National Science Foundation, which is a federal agency ... 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 provide a definitive interpretation for the context artifact shown in table 7. That 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 is not a well-defined concept.
2. The question requires a somewhat arbitrary choice between response categories (excellent vs. good vs. average vs. below average vs. poor).
3. The concept of the job of businessman, for example, is rather imprecise: Does it mean the local grocer or the president of General Motors?
4. The question is one to which people probably give little thought and it has few behavioral consequences in respondents' everyday lives.

PATTERNS OF ASSOCIATION

Our next set of examples concerns a rather different sort of disagreement. In the preceding sections of this chapter, we have been concerned with whether the univariate patterns of response 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 the *patterns of association* between variables measured in different surveys vary systematically. That is, would one 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 bivariate distribution of variables will be undisturbed, even though the marginal (univariate) distributions may vary substantially. As one review of survey research practice recently observed,

> The solution to this problem [of fluctuation in univariate distributions arising from changes in wording] advocated by Davis and other 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—

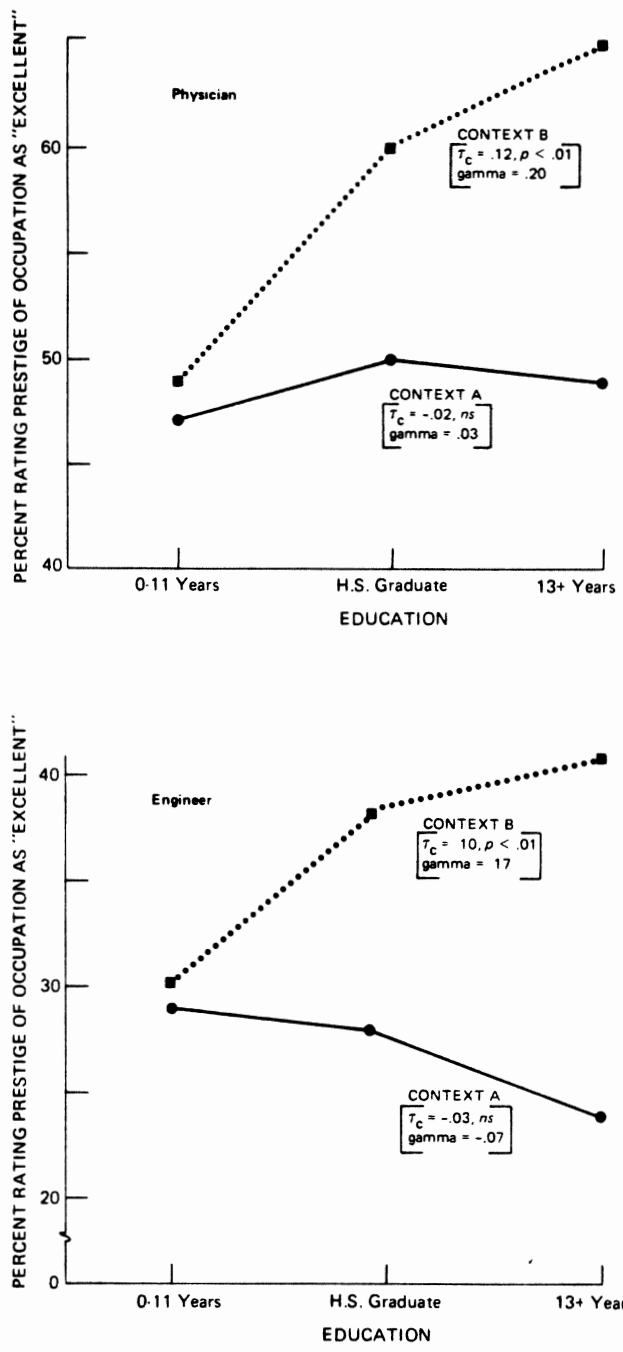
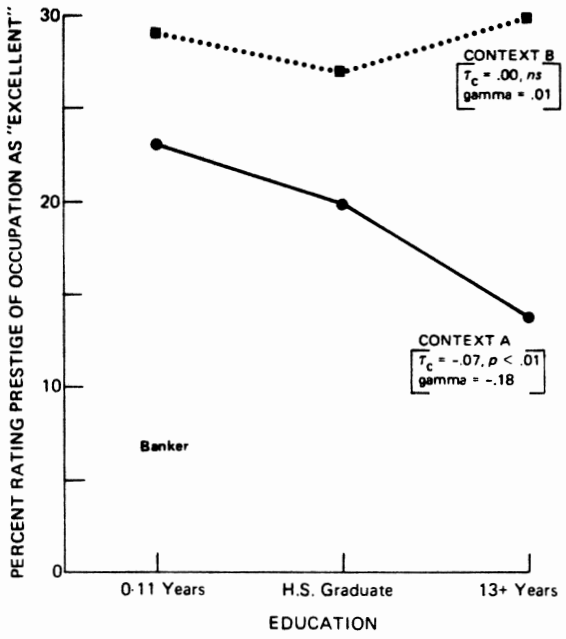


Figure 5. Relationship between ratings of occupational prestige and respondents' educational level for two forms of questionnaire. *Source:* Science Indicators survey, 1976.



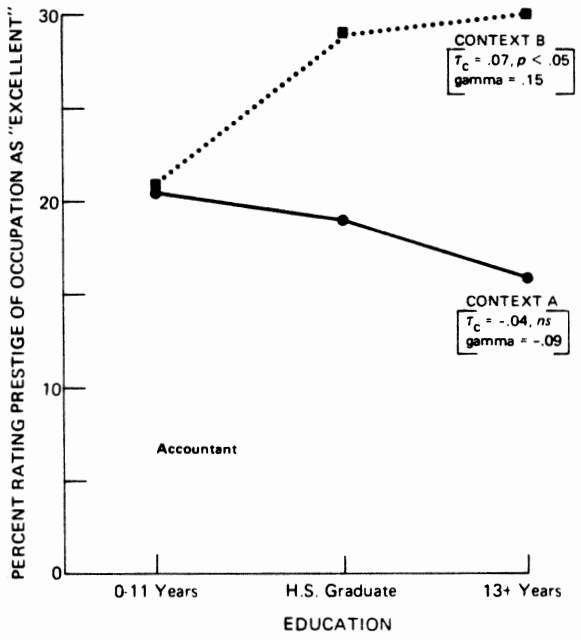


Figure 5. Relationship between ratings of occupational prestige and respondents' educational level for two forms of questionnaire. (Continued)

> and therefore associations among items—are largely immune to this problem. (Schuman and Duncan, 1974, p. 234)

This complacent position has recently been questioned (Schuman and Duncan, 1974; Schuman, 1974; Schuman and Presser, 1977). In their work on the effects of question wording, Duncan and Schuman (1977) concluded that: "[Their research] argues for caution in making the assumption that multivariate patterns of responses will be relatively unaffected . . . even though the univariate response distributions are affected" (p. 9).

Some further information on this question can be gleaned from analysis of the Science Indicators data on occupations. In figure 5 we plot responses to four items from the occupational prestige question by the educational level of the respondents. These results show that the context effects observed in tables 4 and 5 are most pronounced for the highly educated. This, in turn, causes the bivariate patterns of association between respondents' educational levels and their occupational ratings to *vary systematically* between form A and form B. Using ordinal measures of association, we observe modest (median $\gamma = .16$) and generally significant positive correlations between educational level and the likelihood of 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 ($\gamma = -.18$).

Thus, the conclusions one would reach about the relationship of respondents' education and their evaluation of occupations depends 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. Even ignoring the rank information, we would still find evidence for an effect upon the multivariate pattern of response. 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	x^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 or fit were 2; computations assume that clustering effects diminished the effective sample size by one-third.) .

INCOMPLETE EXPLANATIONS: CONTEXT EFFECTS AND PUBLIC CONFIDENCE

In an earlier paper (Turner and Krauss, 1978), large and persistent discrepancies between Harris and HORC 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 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 levels of confidence found by Harris in 1976 and 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 years people responded to the prompt "organized religion," while in other years they were prompted with "how about the people running organized religion."

To provide some experimental evidence upon such context effects, 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 were phrased:

> 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 note . . .
>
> 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; the order of institutions in this table corresponds to their order of presentation in the survey questionnaire.

What can one conclude from these results? At present, our analyses are in a rather preliminary stage; however, 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, the reverse effect was found for the press. When measured after a series of items focusing upon political alienation, confidence in the press rose.

A second conclusion we 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 do note, however, that the NORC experimental manipulation did not fully duplicate the alienation context of the relevant Harris survey. In particular, two questions 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. The omission of these two questions, particularly the one on leadership, does cause some uncertainty in generalizing from these experimental results to the actual measurements made in 1976.

Patterns of Association

Preliminary examination of these data also revealed evidence that the context manipulation had effects upon the patterns of association between confidence

and other variables. Using the three confidence items that showed significant shifts in their univariate distributions, 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 an effect 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 virtually identical assocation with education level (γ = .30) in the two questionnaire forms.

Subsequently, we examined the association between the three confidence measures and responses to the three alienation items (1, 2, and 6) that we judged to be most related (semantically) to confidence in national institutions. Figure 6 presents one example of the basic data. Using log-linear techniques (cf. 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 that excluded the three-way interaction term (CAF), 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 8 provides details of these analyses (we report fits only for the maximally constrained noninteractive model (CA, CF, FA) since this is the appropriate comparison model for testing the null hypothesis of no context effect upon the patterns of association).

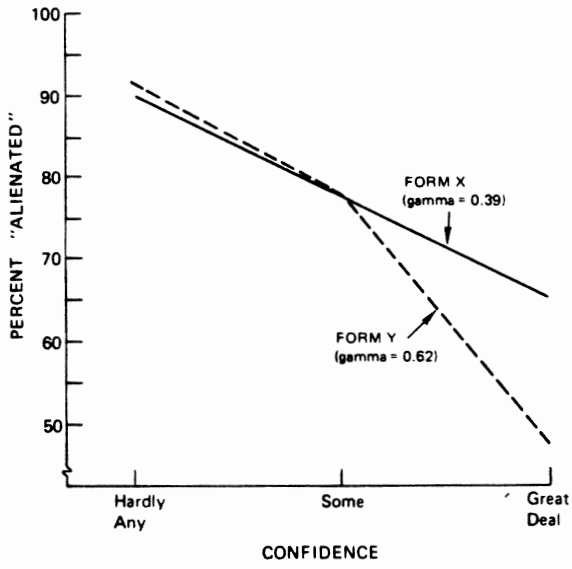


Figure 6. Relationship of confidence in major companies to alienation response for two forms of questionnaire. In form X, the confidence question precedes the alienation items; in form Y, the alienation items precede the confidence question. Alienation item: The rich get richer and the poor get poorer. *Source:* NORC General Social Survey, 1978.

Table 8. Test for Context Effects on Patterns of Association Between Confidence and Alienation

Alienation item	Confidence item	Ordinal association (γ)[a] Form X	Ordinal association (γ)[a] Form Y	Log-linear interaction test[b] χ^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
	Scientific community	.18	.30	1.7	.20
The rich get richer and the poor get poorer.	Major companies	.39	.62	7.4	.007
	Press	.09	.13	0.0	NS
	Scientific community	.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
	Scientific community	.07	.27	4.0	.05

Note: Model fit to response distribution for alienation (A) by confidence (C) by questionnaire context (Q) is a maximally constrained nonsaturated model. In Goodman's notation, it is (CF) (CA) (FA). Failure to fit a model of this type to the data indicates an interaction, that is, that the pattern of association between the variables was not independent of questionnaire context.

NS Not significant.

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

[b]df = 1.

For discussion purposes, let us briefly consider the instance in which this multivariate context effect is strongest. As figure 6 shows, we found 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. $\gamma = 0.39$). An examination of the γ coefficients shown in table 8 revealed that this particular relationship holds true in seven of the nine other comparisons. The two exceptions involved reversals of trivial magnitude.

Although our analyses are still at a preliminary stage and the experiment itself imperfectly replicated the actual context variation, we have nonetheless observed significant effects upon both univariate and bivariate response distributions. However, these effects were neither so pervasive nor so overwhelming in magnitude as to provide a *complete* explanation for the discrepancies observed in the 1972-1977 Harris and NORC confidence series. Clearly, many aspects of the behavior of these disagreeable series remain to be understood.

DISCUSSION AND CONCLUSIONS

This chapter was intended to raise many questions but only to hint at answers. Our collection of disagreeable examples was selected to explore some prelimi-

nary hypotheses concerning the types of subjective indicators that are more (and less) vulnerable to artifacts of measurements. Thus, we have tried to assess the degree to which our hypothesized typology was confirmed or disconfirmed by each example.

In addition to providing some preliminary conceptual organization, we hope that the preceding evidence focuses 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 common and complacent apologia for inconsistencies in survey estimates of subjective phenomena.

First, because of the range of examples presented here and elsewhere (e.g., Cowan et al., 1978; Gibson et al., 1978; Smith, 1978; Turner and Krauss, 1978), one cannot ascribe these discrepancies to the deficient practices of any particular survey research organization. Such a position would be both unfair and unfaithful to the observed facts. We have observed both discrepancies and consistencies in comparisons involving estimates made by a wide range of research organizations.

Second, both our analyses and those recently undertaken by Duncan and Schuman (1977) and Schuman and Presser (1977) indicate that 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 on multivariate patterns of association between variables will be resistant to the anomalies encountered in the analysis of univariate distributions.

Last, it appears that no single explanation is likely to be adequate to explain all the observed discrepancies. Experimental data on artifacts in the confidence time series suggest that context, for example, is probably only a partial explanation for the discrepancies in these time series. Various other sources of error (e.g., interviewer effects) need to be considered.

There is much to be learned about the reliable measurement of subjective phenomena. But we need not be discouraged by our confessions of ignorance. New knowledge is sought when accepted truths are found wanting. At present, reality has belied our expectations about the behavior of these data. Our evidence suggests the need for coordinated research designed to improve our understanding of the validity and reliability of survey-based estimates of subjective phenomena. The growing importance of such data in policymaking and the various social indicators enterprises makes such research imperative.

This research is likely to require coordinated studies across research organizations of the error structure of our survey measurements. While some appropriate work can be undertaken by individual investigators or solitary research organizations, there is a clear need to demonstrate the validity and reliability of

comparisons involving data from diverse governmental and nongovernmental sources. Collaborative tests conducted across many survey research organizations would seem an appropriate strategy for calibrating measurement procedures and improving the reliability of estimates of subjective phenomena. In this regard, the procedures developed to assess analytical measurements made by different laboratories in the physical and chemical sciences may offer useful guidance (cf. Youden, 1975; Steiner, 1975; Boffey, 1975).

The cost and difficulty of improving our understanding of the problems affecting survey estimates of subjective phenomena should not be underestimated. However, those who take seriously the tasks of social reporting and the monitoring of changes in the subjective states of the population would seem to have little choice. Such research needs to be done.

Vulnerable Indicators

At the outset, 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 ambiguous in concept and had little importance for the everyday life of respondents and whose measurement involved a choice between relatively amorphous response categories.

This hypothesis predicts that survey measurements of nonsubjective phenomena, such as chronological age, should be relatively invulnerable to artifacts of measurement (e.g., effects of variant question wordings, survey context). Davis (1976) has studied 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. These estimates do, in fact, show good, although not perfect, consistency with one another and with independent estimates made by the Census Bureau and surveys conducted by NORC.

For subjective phenomena, our analysis of eight pairs of estimates of women's fertility expectations revealed a pattern of consistency, which was within the range expected on the basis of sampling fluctuations. Expectations of childbearing, while clearly a subjective phenomenon, are predicted to be less vulnerable to measurement artifacts by our hypothesis. This prediction follows from the fact that the concept of childbearing is unambiguous, the response categories have a clear meaning, and the question itself is directly relevant to the everyday life of the respondents (married women of childbearing age).

The discrepancies observed in our other analyses involved questions whose character is consistent with our preliminary hypotheses. In table 9, we summarize these comparisons together with those presented in three other recent publications (Turner and Krauss, 1978; Smith, 1978; Kalton et al., 1978). We have included only recent studies because the earlier literature was reviewed by Sudman and Bradburn (1974) in their book, *Response Errors in Surveys*. Their

. Summary of Comparisons of Survey Estimates of Subjective Phenomena

Topic	Response categories	Type of comparison[a]	Comparisons	Significant discrepancies	Maximum discrepancy[b]	
					Raw percentage	z score
fidence in people	Great deal of confidence, only some confidence, hardly any confidence	Different houses[c]	27	18	16	3.6
ıing national tutions		Context experiment	13	3	7	3.3
uation of amount of eral income tax	Too high, about right, too low	Context experiment[c]	1	1	14	5.9
eral happiness	Very happy, pretty happy, not too happy	Different houses[d]	2	1	15	6.2
		Same house/different surveys[e]	2	1	11	5.5
uation of prestige ccupations	Excellent, good, average, below average, poor	Context experiment	10	7	15	5.9
uation of contem- ıry driving standards	Lower than they used to be, higher than they used to be, about the same	Context experiment[f]	2	1	7	3.5
luation of traffic e	Becoming noisier, less noisy, or about the same	Context experiment[f]	2	1	6	2.4
luation of need for ttime truck veries to retail es	Yes, no[g]	Context experiment[f]	2	1	9	≈ 2.8
luation of spending ıational problems	Too much, about right, too little	Different houses[h]	10	5	13	5.5
anthropy I: people to take advantage ou	Try to take advantage if they got a chance, try to be fair	Different houses[h]	2	0	4	1.8
anthropy II: most ime people try to ıelpful	Try to be helpful, mostly just looking out for themselves	Different houses[h]	2	1	6	2.5
ınthropy III: most ple can be trusted	Can be trusted, can't be too careful in dealing with people	Different houses[h]	2	1	7	3.3
ırts' treatment of ninals	Deal too harshly or not harshly enough with criminals	Different houses[h]	1	0	1	0.6
ı or women better ed for politics	Men better suited, equally suited, women better suited	Different houses[h]	1	0	6	1.8
e for woman for ident	Yes, no	Different houses[h]	1	1	5	2.2
dren safe bicycling ocal area	Yes, no	Context experiment[f]	2	0	2	≈ 0.6
requiring police nit to purchase gun	Favor, oppose	Different houses[h]	3	0	2	0.6
port for legaliza- of marijuana	Yes, no	Different houses[h]	1	0	3	1.5
tical party with ch one identifies	Republican, Democrat, independent, other	Different houses[h]	3	1	4	3.2
r death penalty urder	Yes, no	Different houses[h]	3	0	3	1.5
number of ren for family	0, 1, 2, 3, . . .	Different houses[h]	1	0	1	0.6
cted number of ren	0, 1, 2, 3, . . .	Different houses[h]	8	1	9	2.0

Comparisons are restricted to those cited in four recent reviews (Kalton et al., 1978; Smith, 1978; Turner and Krauss, 1978; this chapter). Only ons involving measures of subjective phenomena are included; behavioral measures (e.g., voting) and demographic estimates are excluded. ons involving variant question wording have been excluded—except where noted otherwise. All comparisons made within 5 months of each other.

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

epancies are presented as absolute difference between percent of population giving selected response (raw percentage P), and the z score for the discrepancy,

$$z = \frac{P_1 - P_2}{\hat{\sigma}_{P_1 - P_2}}$$

$_1 - P_2$ is standard error for expected value of difference between two indpendent estimates, discrepancies with z scores greater than 2.0 can be d reliable (i.e., $p < .05$).

nputing these statistics, we have used standard error estimates from the source publications, if available. When data were derived from clustered nd no estimates of sampling error were available in the source publication, we computed estimates based on the formula for a simple random ut we employed deflated sample sizes ($N' = 0.66N$) to allow for design inefficiencies.

er and Krauss (1978).

parison involves a modest difference in question wording, see page 44.

parisons were made by pooling estimates from 2 months of NORC Continuous National Survey closest to dates of General Social Surveys.

n et al. (1978).

ce does not quote explicit response categories; these are implied by text of question.

h (1978).

Surveys of Subjective Phenomena: A Work

conclusions, however, are somewhat at variance with those that have emerged from later work.

The summary presented in table 9 shows a rough correspondence with the typology of our preliminary hypotheses. Thus, we find relatively more significant discrepancies for items 1-14. These items 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 inquires about topics that have been the subject of considerable public discussion (e.g., capital punishment and the legislation of marijuana) or that are connected with specific behaviors that have tangible behavioral components (e.g., voting and party registration or 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 that, we suspect, would not be subject to considerable public discussion as phrased in these questions, that is, Do I have confidence in the people running major companies? What is the prestige of accountants? Are we spending too much on science? Over one-half of the comparisons (42 of 82) involving this group of questions produced significant discrepancies.

Although we believe that a case can and has been made for the typology we hypothesized, table 9 is not without its counterexamples. Voting for a woman president (yes or no) is a concrete action and related issues have been the subject of considerable public discussion. Yet, in the one instance where a comparison was possible, we find a modest (4 percent) but reliable difference between estimates derived from independent surveys using this question. Thus, while the evidence is generally compatible with our hypotheses, the correspondence between our typology and the available data is less than perfect.

Other evidence supports our typology of vulnerable indicators. In an unpublished experiment, Duncan and Schuman (1977) found context-induced variations in responses to five (of seven) questions. These questions were generally consistent with our typology; they measured, for example, 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.

In addition, recent studies (Cowan et al., 1978; Gibson et al., 1978) of survey-based estimates of crime victimization have revealed that substantial and significant variations in respondents' reports of crime were induced by differences in questionnaire contexts. While the National Crime Surveys (NCS) incorporate rather specific descriptions in their questions, for example,

> Did anyone beat you up, attack you, or hit you with something, such as a rock or bottle (In the last 6 months)?

crime victimization rates for each of the 13 cities surveyed by NCS show a strong rise in crime reporting when measurements of victimization are made after a series of attitudinal questions on crime and fear of crime. Gibson et al. (1978) report that this measurement artifact produced a relative increase of 12 percent in the rate of property crime and 21 percent in the rate of personal crime.

Examination of the NCS questionnaire suggests that this result fits within the rubric of our typology. The victimization questionnaire includes, for example, the following items:

> 42. Did anyone try to attack you in some other way (other than any incidents already mentioned)?
> 46. Did you find any evidence that someone attempted to steal something that belonged to you (other than any incidents already mentioned)?

The concepts of *attempted* theft and attempted attack, in these questions, are somewhat amorphous. Our hypotheses would predict that estimates derived from these questions would be more vulnerable to artifact than those derived from questions on the actual incidence of theft and physical assault. We also suspect that these items bear a major responsibility for the observed variations in crime victimization rates (remember, *attempted* theft or assault is also a crime). However, even questions on actual assault are not without their ambiguities, for example, when does a friendly poke become a hostile blow?

Analysis of the NCS data by Cowan et al. (1978) provides some confirmation of our predictions. For example, simple assault and aggravated assault show little evidence of context-induced artifacts (z scores for difference in rates: 0.3 and

0.2). However, attempted assault with a weapon and attempted simple assault show much stronger evidence of context-induced artifacts (z = 1.8 and z = 2.8). Similarly, robbery estimates show less evidence of artifacts than attempted robbery (z = 0.2 vs. z = 1.2), and estimates of personal larcencies involving actual contact between victim and thief are less affected by context than estimates of larcenies involving no personal contact (z = 0.3 vs. z = 3.4).

FUNDAMENTALS AND FUTURE DIRECTIONS

We think it would be a mistake to view the problems presented by these disagreeable examples as strictly methodological. Similarly, we do not think that remedies should be sought in narrowly conceived methodological research.

We suggest that there is a need for a reconsideration of the psychological assumptions that underlie the practice of survey research. We do not have a particular agenda of research to propose in this area, rather we believe that a fundamental reconsideration of the psychological foundations of survey research ought to be encouraged, and we applaud those independent initiatives that have recently emerged (e.g., Nisbett and Wilson, 1977; Wilson and Nisbett, 1978; Fischoff et al., 1979).

It should be obvious that there is a fundamental relationship between psychological concerns and the practice of survey research. While this is doubtlessly a truism, it is often ignored.

The most fundamental phenomena of survey research are quintessentially psychological in character. They arise from a complex interpersonal exchange, they embody (or are contaminated by, if you wish) 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 and the demand characteristics of the interview.

In this regard, we 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 encoders of the subjective reality of respondents. It is a bit odd that as social scientists we must adopt such a narrow view of the social realities involved in our own work.

The burden of the observed anomalies should prompt a reconsideration of the psychological foundations of survey research. The foregoing examples are indicative of the deficient state of our present knowledge, and we hasten to note that such topics have not been the subject of particularly active research in the last decade. We doubt that there are any instant solutions. However, it also seems clear that complacency will not suffice.

REFERENCES

Abrams, M. 1973. "Subjective Social Indicators." In United Kingdom, Central Statistical Office, *Social Trends: 1973.* London: HMSO.

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

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

Blau, P., and O. D. Duncan. 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,* February 24, 1975, 1.

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

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

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

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

Davis, J. 1975a. "Communism, Conformity, Cohorts and Categories: American Tolerance in 1954 and 1972-3." *American Journal of Sociology,* 81: 491-513.

____. 1975b. "Does Economic Growth Improve the Human Lot? Yes Indeed About .0005 Per Year." Paper presented to the International Conference on Subjective Indicators of the Quality of Life, Cambridge, England.

____. 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 Socioeconomic Index for All Occupations." In A. Reiss, ed., *Occupation and Social Status.* New York: Free Press.

____. 1972. "Federal Statistics, Non-Federal Statisticians." *Proceedings of the American Statistical Association (Social Statistics Section).* 152.

____. 1979. "Indicators of Sex Typing." *American Journal of Sociology* 85: 251-260.

Duncan, O. D., and H. Schuman. 1977. *An Experiment on Order and Wording of Questions.* Unpublished manuscript, Department of Sociology, University of Arizona.

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.

Fischoff, B., P. Slovic, and S. Lichtenstein. 1979. "Knowing What You Want to Know: Measuring Labile Values." In T. Wallsten, ed., *Cognitive Processes in Choice and Decision Behavior.* Hillsdale, N.J.: Earlbaum.

Freedman, R., A. Hermalin, and M. Chang. 1975. "Do Statements About

Expected Family Size Predict Fertility? The Case of Taiwan, 1967-1970." *Demography* 12: 407-416.

Gibson, C., G. Shapiro, L. Murphy, and G. Stanko. 1978. "Interaction of Survey Questions as It Relates to Interviewer-Respondent Bias." Paper presented at the 138th Annual Meeting of the American Statistical Association, San Diego, August 14-17.

Goldberg, D., H. Sharp, and R. Freedman. 1959. "The Stability and Reliability of Expected Family Size Data." *Millbank Memorial Fund Quarterly* 37: 369-385.

Goldfield, E., A. Turner, C. Cowan, and J. Scott. 1977. "Privacy and Confidentiality as Factors in Survey Response." *Proceedings of the American Statistical Association (Social Statistics Section).* 1-11.

Goodman, L. 1971. "The Analysis of Multidimensional Contingency Tables." *Techonometrics* 13: 33-61.

____. 1972. "A General Model for the Analysis of Surveys." *American Journal of Sociology* 77: 1035-1086.

Gurin, G., J. Veroff, and S. Feld. 1960. *Americans View Their Mental Health.* New York: Basic.

Hall, J., and D. Jones. 1950. "The Social Grading of Occupations." *British Journal of Sociology* 1: 31-55.

Ho, C. Y., R. W. Powell, and P. E. Liley. 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 Academy of Sciences, *Environmental Monitoring*, vol. 4a, Washington, D.C.: National Academy of Sciences, National Research Council.

Kalton, G., M. Collins, and L. Brook. 1978. "Experiments in Wording Opinion Questions." *Applied Statistics* 27: 149-161.

Kish, L. 1965. *Survey Sampling,* 2d ed. New York: Wiley.

Kraut, A., I. Wolfson, and A. Rothenberg. 1975. "Some Effects of Position on Opinion Survey Items." *Journal of Applied Psychology* 60: 774-776.

Martin, E. 1978. "Trends in Victimization: Problems of Measurement." Paper presented at the 86th Annual Meeting, American Psychological Association, Toronto, August 28-September 1.

____. 1981. "A Critique of Replication Studies of Social Change: Problems in Monitoring Change." In P. Rossi and J. Wright, eds., *Handbook of Survey Research.* New York: Academic Press (in press).

Mason, K., J. Czaijka, and S. Arber. 1976. "Changes in U.S. Women's Sex Role Attitudes." *American Journal of Sociology* 41: 573-598.

National Academy of Sciences, Committee on National Statistics-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.: U.S. Government Printing Office.

____. 1975. *Science Indicators: 1974.* Washington, D.C.: U.S. Government Printing Office.

____. 1977. *Science Indicators: 1976.* Washington, D.C.: U.S. Government Printing Office.

Nisbett, R., and T. Wilson. 1977. "Telling More Than We Can Know: Verbal Reports on Mental Processes." *Psychological Review* 84: 231-259.

Rich, R. 1975. *An Investigation of Information Gathering in Seven Federal Bureaucracies: A Case Study of the Continuous National Survey.* Unpublished doctoral 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 (England), April 1974.

Schuman, H., and O. D. Duncan. 1974. "Questions About Attitude Survey Questions." *Sociological Methodology: 1973-4.* San Francisco, Jossey-Bass.

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

Schuman, H., and S. Presser. 1977. "Question Wording as an Independent Variable in Survey Analysis." *Sociological Methods and Research* 6: 151-170.

Sewell, W., and R. Hauser. 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.: U.S. Government Printing Office.

Smith, T., 1978. "In Search of House Effects: A Comparison of Responses to Various Questions by Different Survey Organizations." *Public Opinion Quarterly* 42: 443-463.

____. 1979. "Happiness: Time Trends, Seasonal Variations, Intersurvey Differences, and Other Mysteries. *Social Psychology Quarterly* 42: 18-30.

Staines, G., and R. Quinn. 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.* Washington, D.C.: Association of Official Analytical Chemists.

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

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

Turner, C., and E. Krauss. 1978. "Fallible Indicators of the Subjective State of the Nation." *American Psychologist* 33: 456-470.

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. 1973. *Social Indicators, 1973.* Washington, D.C.: U.S. Government Printing Office.

____. 1977. *Social Indicators, 1976.* Washington, D.C.: U.S. Government Printing Office.

Waksberg, J. 1975. "How Good Are Survey Statistics?" *Proceedings of the American Statistical Association (Social Statistics Section)*, 26-27.

Wilson, F., and L. Bumpass. 1973. "The Prediction of Fertility Among Catholics." *Demography* 10: 591-597.

Wilson, T., and R. Nisbett. 1978. "The Accuracy of Verbal Reports About the Effects of Stimuli on Evaluations and Behavior." *Social Psychology* 41: 118-131.

Youden, W. 1975. *Statistical Techniques for Collaborative Tests.* Washington, D.C.: Association of Official Analytical Chemists.

Youden, W., and E. Steiner, eds. 1975. *Statistical Manual of the Association of Official Analytical Chemists.* Washington, D.C.: Association of Official Analytical Chemists.

Irregularities in Survey Data

Angus Campbell
Institute for Social Research
University of Michigan

It is always difficult to evaluate the kind of analysis Dr. Turner presents. Do we take his examples of irregularity as isolated incidences having only case history significance, or do they deserve a broader interpretation? Turner finds these discrepancies to have weighty implications, and his chapter gives the impression that the reality of survey data is considerably seamier than our innocent expectations.

Turner describes his analysis as being at "a preliminary stage," and one cannot properly complain that he has not done more than he intended. I am concerned, however, that an unsophisticated reader will take his "disagreeable examples" as a general indictment of the quality of survey data, and I regret that the language of the article rather invites the reader to that conclusion. I feel confident that by a judicious selection of examples I could present a set of data that would make survey findings appear to be as solid as Gibraltar, but it would have the same basic weakness as Turner's piece. Generalities cannot be derived from isolated and purposefully selected incidents.

Turning to the author's first example of where surveys disagree, we find a very curious selection of data. The argument revolves around figure 1, Trends in Self-reported Happiness, 1971-1973. There appears in the *Social Psychology Quarterly* in March 1979, 2 years before the present publication, an article by Tom W. Smith entitled "Happiness: Time Trends, Seasonal Variations, Inter-survey Differences, and Other Mysteries." It deals with precisely the same data that Turner reviews but extends them in time from 1971 to 1977. Smith presents a more extensive review of the context problem and also shows a rather persistent seasonal variation. When these effects are removed, it appears that the NORC and SRC data show a generally similar pattern over the period 1971-1977, although the NORC percentages of people reporting themselves very happy are consistently a little higher than those of the SRC. The reader who sees only figure 1 in Turner's article would surely conclude that the data of the two organizations were heading off in opposite directions with no relationship

to each other. I hope that anyone interested in the reliability of happiness scores will look up Tom Smith's article.

I would also point out to anyone with such interests that the happiness question is one of the least reliable measures we have for the assessment of psychological well-being. It has all the weaknesses of a single-item measure. It persists in current research because it is the only measure of its kind that has any time depth. George Gallup began asking a question about personal happiness in 1946, and the precise wording that both NORC and SRC have been using originated in 1957. We have so few trend data in this field that we cannot afford to throw anything away, even if it is less reliable than the multi-item measures we now depend on.

One cannot argue with Turner's implication that NORC was ill-advised to place a happiness-in-marriage question directly prior to its question on general life happiness. He demonstrates very effectively, as Smith does in greater detail, that it produced a context effect. I am not prepared, however, to accept his suggestion that happiness is such a "notably amorphous" concept that it cannot be measured reliably. It must be remembered that the experience of happiness is subject to variation resulting from changing external circumstances. Even if it could be measured without error, there would be variation from one measure to the next. In 1978 the SRC reinterviewed 694 people we had interviewed in a national sample in 1971. The correlation of the answers to the standard happiness question over the 7-year period was .36, but the correlation between the answers to an eight-item measure of general affect used in the two surveys was .51. Considering the number of life-shaking events that must have happened to these people over these years, I find the degree of constancy shown in the latter measure to be rather impressive.

Turner's second example subjects NORC to another test from which it emerges quite handsomely. The striking fact about figure 3 is the stunning stability in year-to-year Census reports of number of women expecting no more children. Would that we always had samples of 4,000 to follow change over time. I am surprised that Turner did not raise the question of why the two sides of figure 3 are so similar. Considering the fact that the respondent was asked first, "Do you expect to have any more children?" It is obvious that the proportion who are not ever going to have more children must be the same proportion who are not going to have any in the next 5 years. Indeed, although table 3 does not show contingency instructions, one would assume that questions 2 and 3 were not asked of women who answered question 1 by saying they did not expect any more children. Figure 3*b* is in fact superfluous.

The section "Science and the Public" deals with two rather dramatic examples of context effects in surveys done for the National Science Board. They are impressive and they clearly resulted in the waste of public money. Turner approaches this argument by setting up what appears to me to be a preposterous straw man to the effect that "the prevailing wisdom among survey research-

ers" tells us that question position has no effect on question responses. To support this proposition he presents the quotation from Sudman and Bradburn (1974) on response effects. This quote is certainly very delicately selected; in the next paragraph in their book the authors write, "The findings reported here confirm the analysis of a limited number of studies focusing on question order effects (Bradburn and Mason). That analysis failed to show any consistent order effects, although individual studies did report significant order effects.... Considerably more research will have to be done before we can formulate any theory on position effects."

I don't pretend to represent the prevailing wisdom, but I would think that most survey practitioners with any sophistication about interview construction have learned to be wary about question order. They are not always able to predict the effect of question placement and are sometimes surprised that it is larger or smaller than they expected. But certainly if one is setting up a series of studies to measure trends where the integrity of the absolute values obtained is critical, one makes every effort to keep the question order precisely the same from one wave to the next. Any departure invites the kind of discrepancy that appears in this section of Turner's chapter. I am not sure what theory would have given us any basis to predict that intervening questions on hamburgers and litter would influence estimates of occupational prestige, but it would not have taken a lot of imagination to guess that compelling respondents to consider how effective they expected science and technology to be in solving current problems might influence their views of how much money ought to go into these efforts. The only thing that surprises me about this presentation is how poorly the National Science Board apparently is served by the people who advise it on survey research.

Turner returns again at the end of this section to his argument that context influence is greatest in questions dealing with poorly defined concepts. This hypothesis is not unreasonable to be sure, but one observes that the discrepancies in table 4 in regard to such apparently precise concepts as reducing crime, treating drug addiction, improving the safety of automobiles, and the like are generally larger than those in table 6 involving occupational prestige. The critical factor appears to be the content of the first question and how it relates to the content of the next question.

Turner suggests that the response categories (excellent, good, average, etc.) used in the occupational ratings are arbitrary and therefore subject to measurement error. Psychological scales clearly do not have the precision of measures based on cardinal numbers, but with all their faults they often behave with remarkable discipline. Table 5 presents the marginals from a series of studies of public evaluations of the money spent on various public problems. The percentages march across the page in an impressive display of stability. The only one that does not records the growing support for military expenditures, which Congress now finds irresistible. The labels "too little, about right, and too much" do not appear less arbitrary or ill-defined than excellent, good, average, and the like.

In "Patterns of Association," Turner presents an interesting example of contextual influence not only on marginal distributions but on correlations of these distributions with an outside variable. College-educated respondents were clearly more attentive to the questions they were asked than the less educated respondents, many of whom, we may assume, were expressing what Philip Converse calls nonattitudes. One would have to guess that there is a great deal of random variation in these latter responses, which is generally offsetting, and that the standard deviation of these responses is larger than that of the college graduates even though their mean score is more stable.

It is rather wry that figure 5 shows the patterns it does while Turner is arguing that questions "to which people probably give little thought" are the most susceptible to context influence. It seems quite clear that in this case the respondents who could be expected to be giving the least thought to the question were practically impervious to the influence of the question order. I do not think Turner's argument is totally without merit, but if respondents are presented a series of questions they do not understand, it does not seem likely that one would have much influence on their reactions to the next.

In his 1978 article with Krauss, Turner undertook to explain the rather substantial discrepancies existing in a series of NORC and Harris surveys asking respondents to express their degree of confidence in the leaders of certain national institutions. The data were far from ideal for his purpose; two of the five pairs of surveys were taken at an interval of several months, and there was some variation in the order of questions in the NORC surveys. No simple explanation of the discrepancies in the two sets of data came out of this analysis, and Turner proposed that context effects might account for some part of them. He was particularly taken with the possibility that a set of questions intended to measure political alienation, which preceded the confidence-in-leadership questions in the 1976 Harris survey, had deflated the levels of confidence expressed in that survey.

Perhaps in response to Turner's article, NORC undertook the experiment in 1978 described in "Incomplete Explanations" in his chapter. The impressive fact about the results of this study, as shown in table 7, is the total absence of any sensible pattern in the results that might be attributed to the presence of the alienation questions. Of the 13 responses, 1 seems to have changed clearly beyond the range of chance and in the expected direction. Two others changed enough to suggest an influence, but one changed positively and the other negatively. Despite Turner's belief that this was an instance "where one might anticipate substantial context effects to occur," they did not appear. Perhaps if the two additional Harris questions that NORC omitted had been included, his anticipations might have been more satisfactorily fulfilled, perhaps not. My guess is that if one searched about, one could probably find a number of examples of this kind—the questionnaire appears to be loaded for context effect, but the actual influence is either invisible or illogical. As Sudman and Bradburn concluded some 16 years ago, order effects tend to be rather inconsistent.

Turner tells us in his conclusions that his intention is "to raise many questions but only to hint at answers." If I felt confident that his readers would take his examples merely as hints, I would feel more comfortable with his presentation, but in fact I am not sure that he takes them as such himself. He appears to feel that he has "disposed" of house effects as a contributor to discrepancies in survey data, although his evidence would seem quite inadequate for this achievement. He leaves the impression that context effects are a major contributor to survey variability, although after devoting most of his presentation to them he admits they may be only a partial explanation. He suggests fallibility in measures of subjective phenomena that surely requires a much broader documentation than he has given it. It is probably virtually inevitable that a collection of case studies will give an impression of general truth even though, as in this case, the author insists that he hopes merely to provide "initial hypotheses around which research may be organized."

At various points, Turner speaks of a need for "coordinated research" to improve understanding of the validity and reliability of survey data. This need has been present ever since World War II, and it becomes increasingly pressing as time passes. I regret that he did not offer any specific suggestions on what a program of coordinated research would look like.

I would like very much to see a serious inquiry into the question of the reliability of survey measures of both objective and subjective phenomena. This inquiry would require a broad-scale review of data from identical questions asked of comparable samples over some period of time. Its purpose would not be to find intriguing examples of consistent or inconsistent data but to plot the variability of the total corpus of data, to compare its distribution to what we would expect from sampling error alone, and to identify those influences that contribute to unreliability.

My own inclination would be to begin this research by dividing these data by agency of origin. Turner makes it clear, both in this chapter and in his 1978 paper (Turner and Krauss, 1978), that he does not regard house differences as very significant. I believe the differences in the quality of the data produced by the Census Bureau, for example, and by the poorest of the commercial polls are substantial, and I regret the tendency of the general public and many scholars to assume that all sample surveys are equally reliable. These agencies differ, not only in their basic sampling designs, but in the precision with which they carry out these designs, in their attention to supervision, and in their general insistence on high standards at all stages of the survey process.

We need a broad-gauge study for which the surviving organizations, private and public, make available all of their trend data over the last 20 years, that is, all distributions resulting from identical questions asked of comparable samples over that period of years or some part of it. The analysis of these data would answer several questions. First, it would tell us what the range of variability in survey data actually is rather than what we would predict it to be on the basis of the formula for sampling error. Second, it would reveal the differences in

reliability that may or may not exist in different kinds of data, objective and subjective, for example. And third, it would make it possible for the first time to know whether the money, time, and care that some survey organizations put into full-scale probability sampling produce more reliable data than those produced by less expensive methods. This would be a very worthwhile undertaking, and the Survey Research Center would be pleased to be the first to make its files available.

No one who is experienced in the generation and analysis of survey data can be unaware of the irregularities that occur in these data. We have all been confronted with outliers that defy explanation. Most of us would agree that even after 35 years of postwar data gathering, survey researchers still have a good deal to be modest about. But survey research must be judged not only by its frailties but by its strengths as well, and the record of achievement is obviously substantial. Research on survey methodology is moving forward at various points around the country, and we may hope that in due course these inquiries will provide answers to some of the questions that now confound us.

REFERENCES

Smith, T. 1979. "Happiness: Time Trends, Seasonal Variations, Intersurvey Differences and Other Mysteries." *Social Psychology Quarterly* 42: 18-30.

Sudman, S., and N. Bradburn. 1974. *Response Effects in Surveys: A Review and Synthesis.* Chicago: Aldine.

Turner, C., and E. Krauss. 1978. "Fallible Indicators of the Subjective State of the Nation." *American Psychologist* 33: 456-470.

Patterns of Disagreement: A Reply to Angus Campbell

Charles F. Turner
National Research Council
National Academy of Sciences

It is, of course, a privilege to have the benefit of Dr. Campbell's comments. There is, it seems, fundamental agreement between us concerning the modest and unsystematic nature of our present understanding of the nonsampling components of variability in survey measurements of subjective phenomena. Although we seem to be in agreement on this basic point, there are some disagreements between us.

ON EVIDENCE

Dr. Campbell wishes to dismiss the examples presented in "Surveys of Subjective Phenomena" and elsewhere (Turner and Krauss, 1978) as a mere set of case studies. He implies that by the judicious selection of contrary examples, he might demonstrate that such survey measurements are "as solid as Gibraltar."

Assessing the representativeness of any set of examples is an important scientific labor. In the present instance, it is an admittedly difficult task. It is, nonetheless, unfortunate that Dr. Campbell has chosen to rest his argument upon speculation rather than enriching our discussion by presenting evidence in support of his claim.

"Surveys of Subjective Phenomena" does present, in total, 101 recent examples of replicated measurements (table 9), and it reviews their conformance to a general typology suggested earlier by Turner and Krauss (1978, note 12). Doubtlessly, this compilation does not include every recent instance of replicated measurements. It does, nonetheless, represent a substantial number of them. In fact, it is relatively uncommon for two survey organizations to ask precisely the same question at the same point in time. Indeed, one often finds that allegedly identical questions turn out to be worded differently when one

I would like to thank Elizabeth Martin and Theresa DeMaio for their helpful comments on an earlier draft of this manuscript.

consults the actual survey questionnaires. For example, NORC and SRC's happiness measurements (analyzed by Campbell et al., 1976, p. 26; Andrews and Withey, 1976, p. 319; Rodgers and Converse, 1975, p. 130) actually involved questions that differ slightly in wording (although these authors treat them as equivalent without noting this difference).

Because wording differences sometimes have substantial effects on responses, we included only instances in which identically worded questions were asked within 3 months of each other.[1] With one exception,[2] our compilation did not knowingly exclude any substantial body of recent replications. Being a first step, however, we did not perform an exhaustive search but attempted to incorporate only recent published work readily accessible to us. The 101 replicated measurements presented in table 9 of "Surveys of Subjective Phenomena" include all instances reported in the recent work of Smith (1978), Turner and Krauss (1978), and Kalton et al. (1978). Furthermore, drawing as it does upon Smith's (1978) compilation, our analysis includes every known contemporaneous replication of any of the subjective items asked in the six NORC General Social Surveys conducted between 1972 and 1977. In addition, some replicated measurements that were not incorporated among these 101 examples were discussed elsewhere in the text, for example, Census Bureau research on the National Crime Survey measurements (cf. Cowan et al., 1978).

As for the judiciousness of our selection of examples and Dr. Campbell's fear that we invite readers to conclude that a "general indictment of the quality of survey data" is warranted, I would merely ask readers to recall the second example we chose for extended discussion. This example involved measurements of women's fertility expectations. As Dr. Campbell notes, these measurements are quite well behaved. They were selected for early and extended discussion in order to preclude the misreading of our work as a general indictment of all survey measures of subjective phenomena. Our intent is made explicit in our caveat:

> Lest the reader be misled by our first example, we hasten to note that we do not believe that all survey measurements of subjective phenomena are equally vulnerable to artifactual biases. Rather, we wish to delineate areas in which artifact-induced discrepancies might be expected and the factors likely to cause such misbehavior. (p. 48)

---

[1] In this regard, we also note that while Turner and Krauss (1978), report the results of 45 replications involving five surveys conducted by NORC and five by Louis Harris and Associates, our summary excluded two-fifths of those replications. As Dr. Campbell correctly observes, there was a considerable time interval between the NORC and Harris measurements in some years. In the original review (cf. Turner and Krauss, 1978, pp. 458-459 and table 4), separate analyses were done for replications involving overlapping surveys versus surveys that did not overlap in time. It was found that "considering only the [overlapping] measurements, we found that, on the average, discrepancies during this period are marginally greater than those in [nonoverlapping] years" (p. 459). Nonetheless, the nonoverlapping measurements were not included in table 9.

[2] The only substantial body of replicated measurements that we knowingly excluded was the presidential popularity measurements. Those data were not readily available to us at the time; a comprehensive review of those measurements is presently underway.

In conclusion, I suggest that a dispassionate reading of the text does not support Dr. Campbell's attempt to dismiss the evidence as a judiciously selected set of case studies.

ON SURVEY PRACTICES

Dr. Campbell also makes a considerable point of objecting to the claim that question context is often (inappropriately) presumed to have no effect on measurements. He asserts that "certainly if one is setting up a series of studies to measure trends where the integrity of the absolute values obtained is critical, one makes every effort to keep the question order precisely the same from one wave to the next. Any departure invites the kind of discrepancy that appears in this section [of 'Surveys of Subjective Phenomena']." Perhaps, most survey researchers would acknowledge Dr. Campbell's advice as correct, but in fact, they frequently disregard it. Omnibus surveys designed to track trends frequently change their content so that, in fact, questionnaire content and question order are variable over time. Neither the NORC General Social Survey, the SRC omnibus surveys, nor most commercial surveys (e.g., Harris and Gallup) replicate questionnaires in their entirety. Thus, changes over time are confounded with changes that may occur because of variations in questionnaire content and question order. The fact that survey researchers frequently fail to replicate question context from one survey to the next supports the inference that they do not regard the possible effects as very important.

Explicit statements of this presumption and some empirical evidence in support of it may be found in the work of commercial pollsters (e.g., Clancy and Wachsler, 1971).[3] Moreover, prevailing practice, even in scientific papers, does not ordinarily require the allowance that intersurvey comparisons confound population change with changes in survey questionnaires (or other survey procedures). Thus, one seldom encounters analyses expressing the concerns of Duncan and Evers (1975) that

> The study design does not permit us to measure the influence of the prior question sequence on responses to the woman's work questions, [and thus] we cannot rule out an intersurvey difference in frame of reference. (p. 133)

Indeed, two of the leading researchers on subjective measures of well-being (Andrews and Withey, 1976, p. 319, footnote 6), have themselves commented on the same NORC (GSS) happiness data presented in figure 1 and table 2 of "Surveys of Subjective Phenomena." These authors observed that the NORC data show "a sharp and unexpected rise" in 1973-1975. These authors did not recognize or allow for the potential context artifacts that Dr. Campbell believes to be apparent in the NORC (GSS) series.

---

[3] Anecdotal evidence suggests, however, that a few survey organizations have adopted priority systems for ordering questionnaire topics in order to provide some standardization of question contexts across surveys.

In this regard, I would also refer readers to the distressing experience of the Census Bureau's National Crime Survey (discussed briefly in "Surveys of Subjective Phenomena" and more extensively in Cowan et al., 1978, and Gibson et al., 1978). Despite careful consideration and planning, even the most experienced survey research organizations have sometimes found themselves surprised by large and unexpected variations in survey measurements that were induced by context variations initially thought to be unproblematic.

In summary, I suggest that the real problems in this area are not caused by carelessness or a lack of sophistication but rather our lack of theoretical knowledge about the nonsampling sources of variance in our measurements.

SOME QUIBBLES

Having discussed Dr. Campbell's major reservations, I should like to touch briefly on some minor points:

1. *On fertility estimates.* Dr. Campbell correctly points out that a woman's fertility expectations in *all future years* are not independent of her expectations for the *next 5 years*. At a minimum, knowing that a woman expects no further children, one also knows that she expects to have no children in the next 5 years (barring coding errors or logical inconsistency). However, Dr. Campbell errs when he says that the two measurements are redundant. While there is some overlap, unique information is contained in each measurement. For example, knowing that a woman expects no children in the next 5 years does not rule out the possibility that she expects to have children at a later age. Because each measurement contains some unique information, they are not entirely redundant. For this reason, we chose to present both series.

2. *On house effects.* Variations in surveys measurements arise from variations in survey procedures not from the mere fact that a survey is done by organization X rather than organization Y.[4] The aim of scientific research should be to identify those survey procedures that induce variability in measurements.

   Discussions of house effects (i.e., the residual variability in measurements associated with the organization doing the measurement) are not helpful in attempts to improve survey measurements. Such discussions do not identify the *sources* of measurement variability that affect comparisons of data produced by different houses, and thus, they offer no guidance on how to provide better (across house) standardization of measurements.

---

[4] Except for the special case where the mere auspices of the survey induce variations in response. Slight evidence of such effects was found in surveys conducted for the Committee on National Statistics (National Academy of Sciences, 1979) by the Census Bureau and the Survey Research Center (Michigan). Such effects were found for questions that asked (directly or indirectly) about the organizations themselves (e.g., the trustworthiness of data produced by different types of survey organizations). This special case might be defined as a pure house effect.

3. *On sampling.* Dr. Campbell observes that there is a need for assessing the impact of various sampling strategies on survey measurements. In "Surveys of Subjective Phenomena," we purposely focused attention on measurement variability induced by nonsampling factors. This does not imply that we believe sampling factors induce no variability in measurements or that the failure to draw proper samples is not, itself, a problem. Sampling, however, is an area in which there is a well-developed theory and an understanding among users of these data as to what constitutes a good sample.

   Many surveys draw inadequate samples (see, Bailar and Lanphier, 1978); however, the dangers of such deficiencies are well known, and if care is exercised, such deficiencies need not occur.[5] When deficiencies do occur, it is not for want of an adequate theory of sampling. This is not the case when one is dealing with the variability induced by nonsampling factors. In this regard, our present situation is remarkably similar to that described by Dr. Campbell 35 years ago:

   > At the present time the most highly developed aspect of surveying procedures appears to be the sampling. This is not to imply that all surveys, or even most surveys, employ well designed and unbiased sampling methods. It is true, however, that the science of sampling has reached the point that it can select unbiased samples of known probable error to represent virtually any universe a surveyer is apt to be interested in. In contrast, the interviewing phase of survey technology is by no means as well controlled and it is likely that in many surveys the interviewing error is considerably larger than the sampling error. (Campbell, 1946, p. 65)

4. *On contexts and confidence.* In his comments on the confidence experiment, Dr. Campbell correctly observes that the pattern of results in this 1978 General Social Survey experiment was not what we had predicted. What had been predicted was that all confidence measurements would be depressed when measured after the series of alienation items. What we found was that responses to one question were substantially affected (7.4-percentage-point change) and that two other questions evidenced more modest effects.

   While Dr. Campbell observes no pattern in our results, I would point out that the question showing a clear effect in the predicted direction was the one *immediately* following the alienation items. Contiguity, one suspects, plays an important part in determining when one question affects the responses to another.

---

[5] Furthermore, for those, like Dr. Campbell, who are concerned about the untoward impact of compromised sampling, (e.g., the use of quotas rather than full-probability methods at the household level), there are easily applied checks. Indeed, the checks that have been done suggest that for many questions the use of quotas at the household level does not necessarily yield notably different estimates than full-probability methods (Turner and Krauss, 1978, p. 462; Stephenson, 1979).

5. *On history.* Dr. Campbell implies that my own work on the happiness measurements merely rediscovers things already reported by Smith (1979). In fact, my findings, first presented at the 1978 convention of the American Psychological Association, relate to work begun in 1977[6] and originally discussed at an informal meeting attended by several social scientists in the fall of that year (footnote 5, p. 44).

   Tom Smith and I have worked both independently and collaboratively on this and related topics[7] for the past 2 years. I, too, would recommend that readers consult his writings on this subject. Interested readers will find that Smith's suspicions in regard to the 1972-1974 happiness measurements are the same as mine and that his conclusions regarding seasonal variations in these measurements are considerably more cautious than Dr. Campbell's assertion that these measurements show a "rather persistent seasonal variation." Smith writes, "In sum, the hypothesis that happiness (and conceivably other measures of global well-being) follows a seasonal rhythm is plausible, but not proven" (1979, p. 27).

6. *On the need for broad-gauged studies.* Dr. Campbell urges upon us the need for broad-gauged studies of the reliability of survey measurements. On this point, we are in complete agreement.

   Some important work in this direction was initiated by the American Statistical Association (cf. Bailar and Lanphier, 1978), and other important studies have been recently undertaken by individual investigators (e.g., Schuman and Presser, 1981; Bradburn and Sudman, 1979). In the same vein, the National Science Foundation has recently funded a major review of the uses, reliability, and meaningfulness of survey measurements of subjective phenomena. This work is being undertaken by a panel of statisticians and social scientists chaired by O. D. Duncan, under the auspices of the Committee on National Statistics of the National Research Council (National Academy of Sciences). Hopefully, this and similar ventures will go some way toward meeting the important and mutually agreed upon goal of improving both survey measurements and our understanding of their properties and problems.

IN CONCLUSION: ORDERLINESS AMIDST ERROR

In "Surveys of Subjective Phenomena," we argued: (1) that nonsampling factors, such as question context, can induce substantial variations in survey

---

[6] The history of the present publication apparently is typical of most scientific writing. Garvey's (1979) review suggests that 90 percent of scientific findings appear in informal communications prior to publication. Formal publications, in turn, often lag 4-5 years behind informal communications in most disciplines.

[7] Other research on related topics (e.g., Schuman and Presser, 1981; Duncan and Schuman, 1980) also suggests that nonsampling artifacts may be more problematic than previously assumed and that understanding these artifacts is particularly crucial for those wishing to construct time series of subjective social indicators.

measurements of subjective phenomena, and (2) that certain types of survey questions are particularly vulnerable to such measurement artifacts. For example, based on the evidence we reviewed, it appeared that general or summary questions were more vulnerable to such artifacts than specific questions. If the context in which a question is embedded provides a framework that respondents use in decoding the question's meaning, then this result is intuitively reasonable. General (or vague) questions are more labile in their meaning, and thus, they are more in need of interpretation. Hence, question context may play a relatively greater role in determining responses to such questions.

As part of an ongoing program of secondary analysis and new experimentation, we have subsequently carried out three pilot experiments using the general happiness question and the more specific marital happiness item discussed in "Surveys of Subjective Phenomena." Measurements were made using identially worded questions[8] in surveys by the *Washington Post* (May 1979), the Survey Research Center of Michigan (August 1979) and the National Opinion Research Center (February-March 1980). In each survey the order of question presentation was experimentally manipulated. In (approximately) one-half of the interviews,[9] the general happiness item followed the question on marital happiness (*controlled context*); in the remaining half of the interviews, it followed whatever else was in that section of the survey questionnaire (*uncontrolled context*). Since the three surveys were otherwise completely different, the question immediately preceding the general happiness item varied *across surveys* in the uncontrolled context. So, for example, in the *Washington Post* survey, the general happiness item followed a question on income in the uncontrolled context, while in the SRC survey it followed a series of items on the gas shortage and a question asking marital status and in the NORC survey a series of five questions asking "how much satisfaction" the respondent gets from city or place of residence, hobbies, family life, friendships, health and physical condition.

A parallel variation in measurement context occurred for the marital happiness item. In (approximately) one-half of the interviews it followed the question on general happiness (controlled context), and in the remainder of the interviews, it *preceded* the general happiness item and thus followed anything else in that section of the survey questionnaire (same items noted above).

Figure 1 displays the proportion of respondents saying they were very happy in response to the general and marital happiness question in each experiment.[10]

---

[8] General happiness: 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? 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?

[9] In the SRC and *Post* experiments, the sample division was approximately 50:50; in the NORC experiment, it was approximately 66:33.

[10] All samples include only married respondents since the marital happiness question could not be asked of unmarried respondents. Samples are drawn from the adult (18 years and

(Continued)

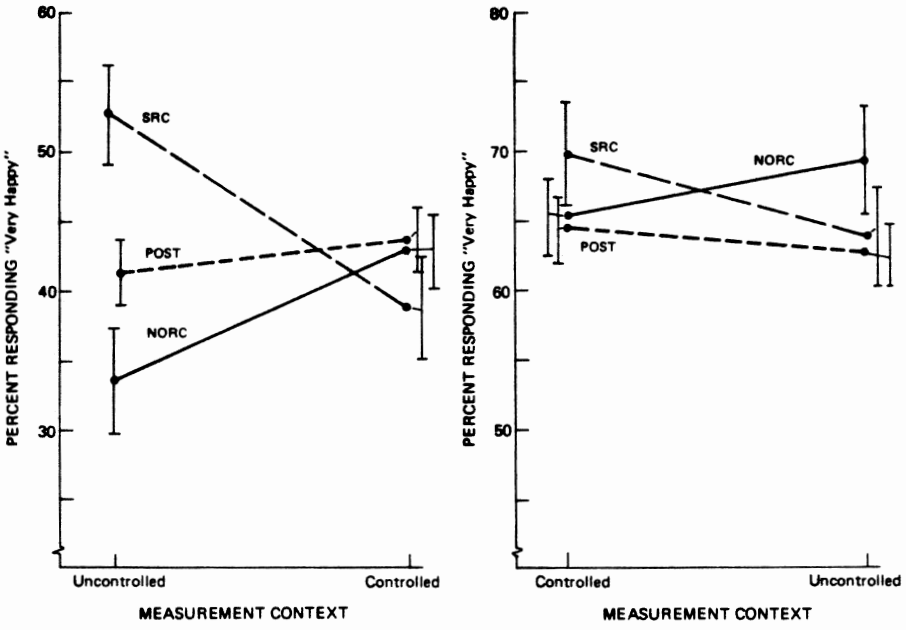


Figure 1. Responses to questions on (*a*) general happiness and (*b*) marital happiness. Error bars demark ±1 standard error for estimates. Data are from surveys conducted by the Survey Research Center, University of Michigan (SRC), the National Opinion Research Center (NORC), and the George Fine Organization for the *Washington Post.* All samples were restricted to married respondents who had telephones in their residences.

While, the results of these experiments are perplexing in some regards, the data do consistently support the hypothesis suggested in "Surveys of Subjective Phenomena." Thus, we note that the more specific question about marital happiness yields equivalent estimates in all three surveys and in both measurement contexts (χ^2 = 10.8; df = 10; p = .37).[11] Secondly, we observe that the

(Continued)

older) population of the continental United States. The SRC and *Post* experiments were done using telephone interviews and random digit-dialing samples; the NORC experiment was done in a face-to-face survey using a multistage area probability sample (see NORC, 1980, for description of sample design). Assignment of respondents to experimental conditions was done at random. Selection of individual respondents in households in the NORC and SRC surveys was done by random selection from all eligible members of the household; in the *Post* survey, selection was from those eligible members of the household who were at home at the time of the initial contact.

All data are unweighted. Due to an error in the preliminary processing of the SRC data, six eligible respondents were eliminated from the tabulation. This error changes none of the SRC results by more than 1 percent.

Data from the NORC experiment include only respondents who reported that they had a telephone in their residence in order to make the sample comparable to those of SRC and the *Washington Post.* (Respondents (circa 3 percent of sample) who refused to give their telephone numbers to the interviewer and thus were not asked the location of their phones were included in this tabulation.)

[11] Likelihood ratio chi-square fit to model 1 (table 1).

general happiness question does yield consistent estimates *but only when the measurement context is controlled.* When the measurement context is left uncontrolled, the measurements vary from 33.5 to 52.7 percent very happy.

Fitting alternative log-linear models (table 1) to the cross-tabulation of happiness response (H) by survey (S) by measurement context (C), we find that a three-way interaction term (HSC) is required to fit the data for the general happiness estimates. In contrast, for the marital happiness measurements, a model (table 1, model 1: H, S, C) positing that the measurements are independent of both context and the survey organization doing the measurements provides an adequate fit to the data ($p = .37$).

While these results fit the pattern hypothesized in "Surveys of Subjective Phenomena," it is important to note that the direction and magnitude of the artifacts found in the general happiness measurements would have been difficult to predict *in advance*. Thus, it is hard to intuit why SRC's measurements in the uncontrolled context are so much higher than those of NORC and the *Post*. Nonetheless, since the different organizations' measurements agree in the controlled context, we can effectively rule out organizational differences in sampling, processing, and so on, as an explanation for the disagreements we have observed.

These results again suggest that our disagreeable survey measurements are not random noise. Rather, disagreements in our survey measurements have identifi-

Table 1. Test of Alternative Models for Behavior of Happiness Measurements

Model	Marginals fit	df	χ^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	(X)
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	(X)

Note: Models were fit using procedures developed by Goodman (1971). χ^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).

X Not applicable.

able causes, and it appears that there are systematic differences between the types of questions that are more (or less) vulnerable to measurement artifacts. If this is true, some improvements in survey practices may follow. At a minimum, in the present case we know that special caution needs to be exercised with general questions on happiness. Replications that do not control the context of such measurements appear to be especially vulnerable to contamination by substantial nonsampling artifacts.

In conclusion, I would suggest that the search for general principles that will allow us to identify and control for the vulnerabilities of our survey measurements are an important part of the work of the next generation of survey researchers. The growing desire to track long term changes in the subjective state of the Nation (proposed elsewhere by Campbell et al., 1976, and Duncan, 1969) requires an improved understanding of the error structure of measurements. This need grows daily with the expansion of the corpus of subjective measurements. The ability to confidently disentangle true change in the population from that induced by changes in our measuring instruments is a prerequisite of reliable inference. On this point, Dr. Campbell and I appear to speak with one voice, and I echo his hope that future research on these questions "will provide answers to some of the questions that now confound us."

REFERENCES

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

Bailar, B., and C. Lanphier. 1978. *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.

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

Campbell, A. 1946. "A Summing-Up." *Journal of Social Issues* 2: 58-67.

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

Clancy, K., and R. Wachsler. 1971. "Positional Effects in Shared Cost Surveys." *Public Opinion Quarterly* 35: 258-265.

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

Duncan, B., and M. Evers. 1975. "Measuring Change in Attitudes Toward Women's Work." In K. Land and S. Spilerman, eds., *Social Indicator Models.* New York: Russell Sage.

Duncan, O. D. 1969. *Toward Social Reporting: Next Steps.* New York: Russell Sage.

Duncan, O. D., and H. Schuman. 1980. "Effects of Question Wording and

Context: An Experiment with Religious Indicators." *Journal of the American Statistical Association* 75: 269-275.

Garvey, W. D. 1979. *Communication: The Essence of Science.* New York: Pergamon Press.

Gibson, C., G. Shapiro, L. Murphy, and G. Stanko. 1978. "Interaction of Survey Questions as It Relates to Interviewer-Respondent Bias." Paper presented at the 138 Annual Meeting of the American Statistical Association, San Diego, August 14-17.

Goodman, L. 1971. "The Analysis of Multidimensional Contingency Tables." *Technometrics* 13: 33-61.

Kalton, G., M. Collins, and L. Brook. 1978. "Experiments in Wording Opinion Questions." *Applied Statistics* 27: 149-161.

National Academy of Sciences, Committee on National Statistics-National Research Council. 1979. *Privacy and Confidentiality as Factors in Survey Response.* Washington, D.C.: National Academy of Sciences.

Rodgers, W., and P. Converse. 1975. "Measures of the Perceived Overall Quality of Life." *Social Indicators Research* 2: 127-152.

Schuman, H., and S. Presser. 1981. *Questions and Answers: Experiments in the Form, Wording, and Context of Survey Questions.* New York: Academic Press.

Smith, T. 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. 1979. "Happiness: Time Trends, Seasonal Variations, Intersurvey Differences and Other Mysteries." *Social Psychology Quarterly* 42: 18-30.

Stephenson, C. B. 1979. "Probability Sampling with Quotas: An Experiment." *Public Opinion Quarterly* 43: 477-496.

Turner, C., and E. Krauss. 1978. Fallible Indicators of the Subjective State of the Nation." *American Psychologist* 33: 456-470.