Timely Artifacts:
A Review of Measurement Variation in the 1972-1989 GSS

Tom W. Smith, NORC, University of Chicago
1155 East 60th St. Chicago, IL 60637

The way to measure change is not to change the measure. By maintaining a constant stimulus over time, one can gauge true change free from distortion due to measurement variation. Conversely, the introduction of measurement variation at any stage in the data collection from sample design, to question wording, to data processing, can distort and invalidate a time series.

Despite the clear necessity and simple principle of replication, constant measurement is often difficult to achieve. First, unintentional variation can easily intrude. This can come from clerical and procedural error, changes in style and method (perhaps due to a turnover in personnel), alterations in "inconsequential" matters, etc. Second, often intentional changes are made. Intentional changes that deviate from strict replication may be adopted for numerous good reasons.[1] They may result from design necessities such as the periodic updating of the sample frame, a desire to maintain consistency with some external standard that has changed (e.g. the switch from 1970 to 1980 Census occupational classifications), competing research goals such as the coverage of new topics or servicing a user community, and/or improvements in data quality.

Third, extra-survey changes may impose variation on surveys despite internal consistency. Extra-survey changes are alterations occurring in society at large which can affect survey results. For example, this might consist of an increase in refusal rates, technological barriers to telephone interviewing (e.g. call screening), or changes in the meaning of words. In a notable, but limited, exception to the strict replication rule survey organizations will sometimes have to change their procedures or level of effort in order to maintain consistent performance standards (e.g. response rate or sample coverage) or to achieve consistency of meaning and validity.

In this paper we will explore the issue of measuring true change by studying in some detail NORC's General Social Surveys (GSS). We will consider 1) how common measurement artifact have been, 2) what the reasons for measurement variation have been, and 3) what can be done to adjust for undesired distortions.

NORC's National Data Program for the Social Sciences has been monitoring change since 1972 with its GSS. The GSS are cross sectional surveys of adults living in households in the United States. The data are collected by in-person interviews. The GSS have been conducted annually from 1972 to 1989 with the exceptions of 1979 and 1981. For more details see Davis and Smith, 1989.

The GSS has striven to faithfully measure true change by the strict replication of measurement procedures such as

question wording, sample universe, coding protocols, and so forth. Several factors have prevented total adherence to the strict replication ideal however. First, certain improvements in the GSS such as the replacement of block quota (BQ) probability sampling with full-probability (FP) sampling and the periodic updating of the sample frame have altered measurement procedures. In general, the GSS has tried to introduce such improvements using a split-sample design under which the old and new procedures are both utilized on random halves in one or more years. This allows for the detection of measurement effects and the calibration and adjustment of the time series to allow for such effects. Second, other basic features of the GSS such as a) methodological experimentation, b) the use of a rotation design to include more variables in the survey, and c) the cross-national research conducted as part of the International Social Survey Program (ISSP) have sometimes led to measurement variation. Finally, a small number of unintentional alterations have occurred.

Because of our switch to a split-ballot design instead of the rotation design, the GSS was recently able to examine for the first time order effects due to rotation variation (i.e. changes in the annual content of GSS due to the operation of rotation of questions, see Smith, 1989). With this source of measurement variation now estimated for the past and largely eliminated for the future, we then carried out a general review of all 600 variables that have been asked at least twice between 1972 and 1989. We perused all of these variables for measurement variations that might have distorted the time series. First, we inspected all variables influenced by measurement variation that had been previously analyzed in the GSS Methodological Reports. Second, we examined all known alterations as documented in Appendix M of the cumulative codebook (Davis and Smith, 1989). Finally, we evaluated each time series for any signs of measurement variations (e.g. the appearance or disappearance of categories, blips, and sudden shifts in distributions). This review identified 119 variables that had been affected to some degree by changes in measurement procedures.[2]

Table 1 lists the 119 affected variables, the type of measurement variation involved (according to the typology outlined in Table 2), the surveys or years altered by the measurement variation, and the type of adjustment that can be made to create a consistent time series.[3] In the table notes the source of the measurement variation is discussed and the steps taken to eliminate or minimize the distortion are indicated.[4] In Appendix 1 the corrected times series are presented along with the SPSSx code needed to make the required revisions.[5]

Table 3 summarizes the causes of the measurement variation detailed in Tables 1 and 2. Changes in screens or filters have been the most common alteration (24.3%). In almost all cases sub-groups that were initially excluded from a series of subordinate questions were later included. For example, on the approval of hitting and the approval of hitting by police items, the

general approval of hitting and hitting by police questions served as screen to the situational hitting questions in early years, but no filtering was applied in later years. Second comes order effects (22.8%), either due to planned, experimental manipulations (12.1%) or other context changes (10.7%), often because of the rotation design. Third in occurrence are changes in coding procedures (22.1%). This most frequently involved the sub-division of cruder codes into more refined categories. For example, income ranges have been sub-divided numerous times as inflation pushed current dollar earnings higher and higher. Fourth, there are differences resulting from sampling improvements (7.1%), either the switch to full-probability sampling or the updating of the sample frame. Fifth, there are a few instances (5.7%) in which wordings have been changed. These primarily concern differences in the network items asked as part of the 1985 and 1987 topical modules. Sixth, other alterations (10.7%) cover such matters as changes in interviewer specifications, physical layout, and mode of administration. Finally, in a number of instances (7.1%) changes in society at large about the meaning and/or categorization of terms have led to measurement changes. The most prominent of these is the inflation-induced changes in the "meaning" of current dollars, while the racial definition of Hispanics is another example. In these cases, changes in meaning led to changes in measurement procedures.[6]

In the vast majority of cases the measurement variation has been the result of intentional changes (Table 4). Most intentional changes have been adopted to improve the GSS, such as the alterations in sampling, the refinements of codes by sub-division, and the loosening or dropping of screens (54.6%). Experiments are the second most frequent source of intentional changes in measurement procedures (13.4%). These largely consist of question order/context comparisons. Other intentional changes (8.4%) result from some early switches in coding conventions, the alteration of the network items used in the 1985 and 1987 topical supplements, and changes to match measurement procedures employed by the Census, American National Election Studies, ISSP, or baseline NORC surveys.

Unintentional changes account for 23.5% of all affected variables. We have counted here context effects related to the rotation design (even though the rotation design itself was intentional) and various small, unplanned variations in coding procedures, wordings on show cards, interviewer specifications, and the like.

Overall, relatively few time series have been seriously distorted by measurement variation (Table 5). Almost 81% have been unaffected by measurement artifacts. For another 11% totally consistent time series can be created because the alterations were designed in such a way to be matchable with earlier procedures. This includes sub-divided codes which can be collapsed into the original

cruder codes, modified screens for which the initial screening rule can be applied to all years, and similar situations. Together this means that undisturbed times series exist for 91.3% of all variables.

For another 6.5% consistent time series can be constructed by applying certain adjustments. For example, items affecting context due to rotation design can be adjusted by using the 1988 and 1989 ballot comparisons (Smith, 1989) or earlier experiments (Smith, 1986) as a standard. Also, the impact on a time series of the addition or deletion of codes can be minimized by appropriate recodes and exclusions.

For the last 2.2% of the variables, notable modifications to the time series are necessitated by the measurement variation. For seven variables we have truncated the time series, either deleting some variant year(s) and/or creating two series instead of one. For another seven variables, all involving variables asked only twice, we find that the variations are too great to consider the readings as comparable and there is no firm basis on which to adjust or reconcile the data points.

Not surprisingly the intentional changes are much more likely to be fully correctable than unintentional changes. About 69% of intentional changes can be so adjusted, while only 8% of the unintentional changes can be.

In many cases the corrected time series differs in only minor degree from the raw trends. Sometimes however major, systematic revisions occur. Table 6 lists three examples.[7] First, the growth in racial tolerance (Support for open housing laws - RACOPEN and disapproval of laws against interracial marriages - RACMAR) is slightly exaggerated by the raw numbers because blacks, who were asked these questions only after 1977, are more supportive of racial integration than whites. Second, the time series on the belief that people are helpful (HELPFUL) becomes much more stable when order effects are removed. Finally, the upward surge in proportion Mormon since 1983 is eliminated when the impact of the adoption of the 1980 sample frame is adjusted for. (Other adjusted series appear in Appendix 1.)

Overall, GSS has succeeded in its mission of monitoring true change free of measurement effects. When changes in measurement procedures have been implemented, steps have routinely been taken to insure the continuation of a consistent time series. As a result, in the vast majority of cases reliable time trend analysis is possible. One must however pay close attention to the measurement variation detailed in the cumulative codebook and GSS Methodological Reports, so that appropriate adjustments can be employed when warranted. Without such care one could end up studying timely artifacts instead of true change.

To maintain a reliable time series and minimize measurement variation one must first try to eliminate unintentional changes by strictly replicating methods and by paying close attention to detail. Second, intentional changes should be avoided whenever possible. When there are compelling reasons to make a

change, the change in method should be benchmarked against the standard method, usually by experimental comparisons. It will often be desirable to repeat the benchmarking experiment over several years to get a more exact reading of the methods effect and see if the relationship is temporally stable.[8]

Third, even assuming that unintentional changes can be minimized through close attention to strict replication and meticulous attention to details and that intentional changes can be avoided whenever possible and benchmarked (e.g. by experimental designs) when necessary, that still leaves the difficult issue of dealing with extra-survey changes. These changes tend to fall into two groups: 1) changes in the effectiveness of existing procedures so that the same procedure and level of effort does not yield the same result and 2) changes in meaning.

For example, if, as it has frequently been alleged, refusal rates have risen over the last several decades, then more effort and/or new procedures may be needed to achieve the same response rate formerly obtained. First, one would have to decide whether maintaining the established procedures and efforts or the same outcome would best represent strict replication and minimize measurement variation. In deciding whether to maintain procedures and/or efforts or to change them to obtain a constant response rate, several matters must be considered. First, one may not be able to compensate for the change in society's willingness to cooperate. Second, the necessary changes may be prohibitive in terms of cost and/or time. Third, even when the methods and resources exist, the necessary changes may obtain a constant response rate, but themselves distort the time series. For example, the widespread expansion of monetary incentives to maintain the response rate may fundamentally redefine the nature of the interviewing process and change the quality and quantity of the information obtained.[9] Ultimately these are empirical questions and only close examination of the data will indicate what mixture of new and old will in fact minimize measurement variation.

The second type of extra-survey changes, changes in meaning, are even harder to deal with. In this case the questions and/or response scales alter their standard meaning over time so the words, but not the meaning, remain constant. In one area, monetary matters, meaning changes are rather common. Except at zero inflation the nominal dollar amounts referred to in questions change their true value (meaning) from year to year. For example, a question asked in 1935 about the Townsend plan of giving $200 a month to each couple over 65 was then an extremely generous proposal. But if asked today it would sound rather miserly. Of course, in the special case of dollar amounts one can standardize by converting into constant dollars. An adjustment that would change the 1935 Townsend allotment into approximately $2,000 per month by the late 1980s (Smith, 1987).

Outside of the special case of monetary references, there

usually is no empirical basis for adjustments for changes in meaning. Fortunately, language tends to change slowly[10] and it appears that over the 50 some years that national surveys have been in collecting information that relatively few questions have been rendered incomparable because of changes in meaning (Smith, 1987). Yet changes in meaning do occur and when they occur it is extremely difficult to compensate for it. For example, in 1954 Gallup asked, "Which American city do you think has the gayest night life?" If asked today, San Francisco would presumably rank well above its fifth place showing. Yet what could one do if one wanted to use the 1954 survey as a baseline? We could rephrase the question to ask about the "liveliest," "most entertaining," "most enjoyable," "most exciting," or some similar term. But how would we know that these could be compared? We might even do an experiment to see if these synonyms produced different results in 1990. If they did, we would have to give up any comparison to 1954, but even if they did not, any comparison to 1954 would be uncertain. (For more examples, see Smith, 1987).

Another example of changes in meaning are the switches during the last 80 years in the the polite and generally used term for Americans of African ancestry from colored to Negro to Black and now perhaps to African American. In this case, surveys have made the change along with society, but without any empirical evidence on its possible impact.

Measuring true change is a difficult task. In general one masters the job by strict replication. But unintentional changes easily occur and intentional changes are often necessary. When justifiable alterations are made, benchmarking and calibration are usually called for. Even more challenging are extra-survey changes, which may necessitate that procedures and/or terms actually be changed in order to maintain consistency or at least to minimize artificial variation. To insure the reliable tracking of true change and to keep methods variation to a minimum, constant vigilance and usually constant measurement are needed.

References

Davis, James A. and Smith, Tom W., General Social Surveys, 1972-1989: Cumulative Codebook. Chicago: NORC, 1989.

Ligon, Ethan, "The Development and Use of a Consistent Income Measure for the General Social Survey," GSS Methodological Report No. 64. Chicago: NORC, 1989.

Smith, Tom W., "The Art of Asking Questions, 1936-1985," Public Opinion Quarterly, 51 (Winter, 1987a), S95-S108.

Smith, Tom W., "Ballot Position: An Analysis of Context Effects Related to Rotation Design," GSS Methodological Report No. 55. Chicago: NORC, August, 1988. Revised November, 1989 with 1989 GSS Appendix.

Smith, Tom W. "Can We Have Any Confidence in Confidence? Revisited," GSS Methodological

Paper No. 1. Published in Denis F. Johnston, ed., Measurement of Subjective Phenomena, Washington, D.C.: GPO, 1981.

Smith, Tom W., "Ordering Context Effects," Paper presented to the Context Effects in Survey Conference, Chicago, July, 1986.

Smith, Tom W., "A Preliminary Analysis of Methodological Experiments on the 1984 GSS," GSS Methodological Report No. 30. Chicago: NORC, 1984.

Stephenson, C. Bruce, "Probability Sampling with Quotas: An Experiment," GSS Methodological Report No. 7. Published in Public Opinion Quarterly, 43 (Winter, 1979), 78-95.

Table 2
Reasons for Changes in GSS Marginals Across Surveys
Other Than True Change

I. Survey-Related Measurement Variation
   A. Wording
      1. Text
      2. Show Card
   B. Screens
   C. Order
      1. Context Experiments
      2. Rotation Related
         a. General
         b. Other
            i) Switches between rotations
            ii) Switches to rotation from permanent
      3. Additions
      4. Deletions
   D. Coding/Response Categories
      1. Subdivisions
      2. Other Alterations
         a. Categories Dropped
         b. Categories Added
         c. Categories Redefined
   E. Sampling
      1. Frame (1960/1970/1980)
      2. Procedure (Block Quota vs. Full Probability)
   F. Other
      1. Layout
      2. Interviewer Specifications
      3. Mode of Administration
      4. Not Certain
II. Non-Survey Measurement Variation
   A. Meaning of Terms
      1. Words
      2. Dollar Amounts
   B. Categorizations

Table 3
Causes of Measurement Variation

Wording	5.7%	(8)
Screens	24.3	(34)
Order		
Experiments	12.1	(17)
Other	10.7	(15)
Coding	22.1	(31)
Sampling	7.1	(10)
Other	10.7	(15)
Meaning	7.1	(10)
		(140)

Table 4

Reasons for Measurement Variation

Intentional		
Experiments	13.4%	(16)
Improvements	54.6	(65)
Other	8.4	(10)
Unintentional	23.5	(28)
		(119)

Note: This table is based on the 119 variables listed in Table 2. When more than one factor affected a variable, the variable was classed according to the factors exerting the greater influence.
Other reasons for intentional changes in the survey instrument are a) the convention of dropping Depends categories (4), b) changes in the topical network module (3), and c) changes to confirm to other surveys (4).

Table 5

Changes to Time Series Due to Measurement Variation

No time series possible	1.1%	(7)
Truncated time series	1.1	(7)
Adjusted time series	6.5	(40)
Adapted time series	10.6	(65)
Unaffected time series	80.7	(497)
		(616)

Endnotes

1. Of course they may also be adopted for various poor reasons as well. Many data collection programs have shown little dedication to the principle of strict replication and have introduced changes with little concern about preserving a time series. For example, 1) the American National Election Study has twice revamped the standard response scale used for its policy preference items and made numerous other changes in wording, 2) the US Census adopted its new ancestry measure without benchmarking against its old parental nativity item, and 3) Harris' confidence in institution indictor has varied the number, order, and description of institutions rated over time (Smith, 1981a).

2. We have tried to be inclusive in our coverage and have included a number of cases where the distortion from the measurement variation has been quite minor and no adjustment to the time series is required. On the other hand we have not listed cases involving known variations when here is no indication of resulting changes in the distribution of items. For example, while all time series are technically affected by shift from BQ to FP sampling between 1974 and 1977 we have listed only the four variables (WRKSTAT, SEX, COOP, and ADULTS) that previous research (Stephenson, 1979) indicated showed significant differences. Changes in measurement procedures for particular variables are listed in Appendix M of the cumulative codebook (Davis and Smith, 1989) and many measurement variations are discussed in depth in the GSS Methodological Reports.
   Future research, especially into rotation related context effects (Smith, 1989), may well suggest additional affected variables.

3. Table 1 is available from the author.

4. In a number of instances the adjustments employed here will not best suit all research purposes. For example, we have, when necessary, sacrificed cases within a year in order to keep more years in the time series. In the cases of racial attitudes that means that we have eliminated blacks from the samples in order to keep in the time series years prior to 1978 when most racial questions were not asked of blacks. Obviously a researcher interested in changes over time in white-black differences on racial matters could not adopt this solution and might either restrict his analysis to the few items that have always been asked of both races or to the period since 1978. Similarly, various recode procedures we used will not be most appropriate for all analyses.

5. Appendix 1 is avaliable from the author. The adjustments applied in Appendix 1 are all single factor adjustments. For example, the adjustments to labor force status (WRKSTAT) are based on the observed difference between BQ and FP samples on this item. It does not also try to adjust for other factors such as the undersample of men in FP samples or the oversample of blacks in 1972. Adjustments that simultaneously take into account multiple factors are not developed here.

6. Cases in which the same wording takes on different meaning over time are highly interesting, but rare occurrences in survey analysis (Smith, 1987). We have noted in Table 1 all cases that we are aware of where a change in meaning led to an alteration in measurement procedures and a resulting shift in distributions. We are aware of one instance involving evaluations of "China" where a meaning shift may have occurred, but no change in measurement procedures has been carried out and no change in distributions can be clearly identified with this shift in meaning. Details available from author.
   On dealing with the problem of changes in real vs. nominal income see Ligon, 1989.

7. Table 6 is available from the author.

8. The GSS has done this with the full-probability and block quota experiments in 1975 and 1976 (Stephenson, 1979) and with the 1984-1989 spending priority comparisons (Smith, 1984). Replication is not always practical however.

9. Or consider the case in which the response rate rises when the level of effort is constant. Under the same outcome standard, one could argue that effort should be lessened to obtain the same response rate as before. In this situation, an alternative approach might be to accept the gain in response rate, but to compare results from the full sample to a sample censored to match the response rate of earlier surveys. If no differences appeared, then one might compare the new and old samples (and possibly justify lowering effort in future surveys). If a difference appeared, them one might use the censored data for time series comparisons.

10. Language tends to be stable if slang and fad words/phrases are avoided.