Editor's Note:

This issue is the first published since the annual meeting of the American Association for Public Opinion Research in May. We thought we could make a contribution by presenting shortened versions of at least some of the many papers that contained useful nuggets worth passing along to those who couldn't attend.

We chose three papers that deal with the very important issue of "form effects"—how the structure of the questionnaire or interview affects response. In this instance all three papers focus on what happens to the distribution of responses as response categories offered to respondents change or as question wording changes. All three are based on experiments in which wordings are systematically altered. Two papers, by Jesse Marquette and by Robert Mason, John E. Carlson and Marti McCracken, show how existence and arrangement of response categories read to respondents can affect the frequency of response to questions. The paper below, by George Bishop and Andrew Smith, reminds us of the existence of a rich "mother-lode" of split-ballot experiments done between 1938 and 1988, but mostly before 1950, by The Gallup Organization to test variations in question wording. Although they have been around for quite a while, they have not yet been analyzed enough to yield generalizations about the effects of different wording. Nevertheless, Bishop and Smith believe that the body of experiments actually address all the issues that survey research scholars have raised about question wording and have the potential of yielding information that can help us decide how to ask our questions in given situations.

GALLUP SPLIT BALLOT EXPERIMENTS

By George Bishop and Andrew Smith

The use of split-ballot experiments to investigate the response effects of variations in the wording and context of survey questions has had a long, but somewhat neglected, history in public opinion research. Though many researchers have long been aware that Gallup and Roper, among others, began to experiment with their questions not long after they founded their polling organizations in the 1930s, few, we suspect, realize just how frequent and extensive these early explorations truly were. Furthermore, only a relative handful of them have ever been reported, most notably *Gauging Public Opinion*, edited by Hadley Cantril and his research associates in the Office of Public Opinion Research at Princeton.[1] There is but minimal awareness of this vast data resource for secondary analysis.

We should also mention here were it not for a conversation with Tom Smith about the old split-ballots done by Gallup and others, we might not have discovered them.

The Gallup Experiments

Our report is based on a "content" analysis of photocopies of the original questionnaires used in 332 Gallup Polls in which split-ballot experiments were conducted between 1938 and 1988.[2] The vast majority were conducted between 1938 and 1949. We can only speculate as to why the sudden decline occurred after 1949. Schuman and Presser have suggested several reasons for the general demise of question-wording experiments which became evident in the early 1950s:

1. That marginal distribution of percentages for an item could be affected by the wording of a question was well-recognized by that time by nearly everyone doing major public opinion surveys.

2. There didn't seem much point to demonstrating over and over that the results of polls could be affected by the way questions are worded.

3. While many came to believe that marginals could be affected by question-wording variations, they assumed that associations or correlations between items would not be so affected—the assumption of "form-resistant correlations."

4. Much of this early experimentation was done in an ad hoc manner designed more to illustrate the pitfalls of question wording than to understand the theoretical reasons for such question effects.[3]

All of these reasons may explain much of the overall decline of question wording experiments during this period, but it would not seem to account for the rather precipitous drop in split-ballots at Gallup after the late 1940s. The only other reason we could think of, other than perhaps a major change in personnel at Gallup that no one seems aware of,[4] is that the failure of the Gallup and other public opinion polls to forecast accurately the 1948 presidential election forced them to focus their resources on ways of improving their *sampling* and *interviewing* procedures.

It is interesting to note here that at the time the Gallup Organization was abandoning the split-ballot, the Institüt für Demoskopie at Allensbach in West Germany was getting going under Elisabeth Noelle-Neumann's leader-

ship and beginning to experiment vigorously with the split-ballot technique, which it has continued to date.[5] So this methodological tradition does persist at Allensbach. The Roper Organization also uses split ballots.

Types of Split-Ballots

We identified a little over 3,000 split-ballots in 332 different polls.[6] On the average, then, there were about 9 split-ballots per survey. A dozen to fifteen or more were not uncommon during the peak years of Gallup's experimentation. Many of the so-called split-ballots were done purely for practical reasons having nothing to do with wanting to know how responses were affected by the way in which a question was worded, and the like. Roughly a fourth of the split-ballots, for example, were used to ask either *different questions on the same subject* (11%) or completely *noncomparable questions* (16%) of various kinds. If we add to that the split-ballots that asked the *same question* on each form, *but about different objects or ideas*(12%), then we have nearly four out of ten variations between forms (39%) done largely for the reason of maximizing the amount of information they could get out of any single survey they decided to split. Still, about 60% of the split-ballots which were designed as "true experiments" with the wording, format, context, or some other aspect of the question. Gallup conducted roughly 250 response order experiments alone, for example, a third of which involved exact replications (e.g., one on political party identification done about 20 times). So, there are more than enough available to test the generalizability of various hypotheses that have been developed in recent years to explain different types of response effects.

The first "known" split-ballot was conducted by Gallup in March 1937.

The question had to do with President Roosevelt's attempt to "correct" or "pack" the Supreme Court (depending on your point of view) by increasing the number of members from nine to twelve. All respondents were first asked if they thought "some kind of change is necessary regarding the Supreme Court," and if yes, "Why?" They were then asked if they were "in favor of President Roosevelt's proposal regarding the Supreme Court." Here they split. On *Form 1*, respondents were asked: "Do you think a majority of the nation's voters *approve* of Roosevelt's plan?" On *Form 2*, they were asked: "Do you think a majority of the nation's voters *disapprove* of Roosevelt's plan?" As it turned out, it made no significant difference.

We found it difficult to categorize the myriad ways in which Gallup had varied the wording of his questions in subsequent experiments. After examining and reexamining hundreds of them, it occurred to us that the great majority of these wording experiments involved a variation in one or both of two underlying dimensions:

1. *Abstractness* and *concreteness*, or *generality* and *specificity*, and

2. *Affective tone* (positive-negative) or *social desirability*.

Abstract-Concrete Dimension. The first example comes from a Gallup Poll conducted in June 1943. On Form K, respondents were asked whether the government should "require every family to put *15% of its income* into war bonds," whereas on Form T they were asked a more concrete question about whether the government should "require every family to put *15cents out of every dollar of its income* into war bonds."

In the second example, drawn from a Gallup Poll in October of 1945, respondents who received Form K were asked if they approved or disapproved of the US making "a loan of $6 billion *to help Russia get back on its feet*"—a somewhat vague justification. In contrast, on Form T they were asked, more specifically, whether they would approve such a loan to Russia "*to repair her war damages, build up industry and raise the standard of living of the Russian people.*" The third example, taken from a poll done in late August and early September 1947, shows two rather broad questions: the first asking respondents whether they have "heard or read about *the trouble which England has been having in getting back on her feet* (Forms K-1 and K-2)"; the second, whether they have "heard or read about England's *financial troubles* (Forms T-1 and T-2)." Though the latter is a rather general question, it is nonetheless somewhat more concrete than the former in that it at least specified the *type* of trouble England was having.

Affective Tone/Social Desirability. One example, from a March 1946 Gallup poll, illustrates a variation in the affective tone of questions created by the use of the terms, "big business" or "union labor" rather than the more neutral designations, "business" or "labor." Another, drawn from a February 1949 poll, recalls the classic experiment about forbidding or allowing speeches against democracy, which Schuman and Presser have used as an illustration of wording tone. Form K respondents were asked if they believed "that labor laws should or should not guarantee (insure) the right of labor unions to strike," whereas Form T interviewees were queried on whether they believed "that labor laws should or should not allow labor unions the right to strike?"

Variations in Both Dimensions

In making one of two question versions more specific or concrete, Gallup sometimes altered, simultaneously,

the affective tone or social desirability of that form of the question. In the first illustration, from a March 1938 poll, the addition of the phrase, "in the poor communities," not only makes the question more specific for the respondent, but also makes more socially desirable a "yes" response on whether "the federal government should give money to states to help local schools?" Another example, from the summer of 1945, also illustrates how, in specifying a particular condition in one form of the question, Gallup would often alter the affective appeal of the item. Form T introduced the name of Spanish dictator Francisco Franco as a circumstance into a question asking whether Spain should be "permitted to become a member of the UNITED NATIONS." On Form K, people were asked the more abstract and more neutral question: "Should Spain become a member of the UNITED NATIONS under its present government?"

More generally, we think the two basic dimensions found to predominate in the Gallup experiments may underlie other types of response effects as well. Introducing a counterargument into a question, for example, not only adds an element of concreteness; it also typically changes the affective meaning of the item. So, too, does making various types of response alternatives explicit, rather than implicit, make a question more concrete or specific and often different in social desirability—leading respondents to retrieve rather different information from their memories.

The Gallup split-ballots of 1938-1949 are a rich legacy for today's researchers interested in wording and form effects. They will sustain further work on just about every hypothesis in the literature on these issues.

Endnotes

[1]Donald Rugg and Hadley Cantril, "The Wording of Questions," pp. 23-50 in H. Cantril (ed.), *Gauging Public Opinion* (Princeton: Princeton University Press, 1944). See also Stanley Payne, *The Art of Asking Questions* (Princeton: Princeton University Press, 1951) and Valerie Tamulonis, "The Effect of Question Variations in Public Opinion Surveys," (Master's Thesis, University of Denver) for analysis of Gallup and NORC experiments conducted in the late 1930s and early to mid-forties, as well as a small number of reports on effects of question wording issued in the mid-1940s by NORC, listed in *NORC Social Research.*

[2]We would like to thank the staff at the Roper Center, particularly Marc Maynard for his diligent efforts in locating the Gallup split-ballots.

[3]Howard Schuman and Stanley Presser, *Questions and Answers in Attitude Surveys* (New York: Academic Press, 1981).

[4]Personal correspondence with Paul Perry, a former Vice Chairman at the Gallup Organization, revealed no major personnel changes or other factors that might account for the sudden fall-off in split-ballots after the late 1940s.

[5]Elisabeth Noelle-Neumann, "Wanted: Rules for Wording Structured Questionnaires," *Public Opinion Quarterly*, 1970-71, Vol. 34, pp. 191-201.

[6]This is actually a somewhat conservative estimate since many of these split-ballots—for example, those on question order and context—involved multiple questions, all of which could be treated as separate potential response effects in any analysis.

George Bishop is professor of political science, and senior research associate, and Andrew Smith is research associate, the Institute for Policy Research, University of Cincinnati.

ONE GALLUP SPLIT BALLOT SURVEY OF MARCH 15-20, 1946

Form T

Question: At the present time, which do you think has MORE influence on the laws passed by Congress—business or labor?

33% Business
51% Labor
16% No Opinion

Form K

Question: At the present time, which do you think has MORE influence on the laws passed by Congress—big business or union labor?

34% Business
50% Labor
16% No Opinion