16

The Popularity of Presidents: 1963–80

Michael B. MacKuen and Charles F. Turner

For four decades, polls have sought to measure the popularity of incumbent presidents, and the results of these polls have been widely reported in the media. It is claimed that "amidst the avalanche of polling data filling our newspapers and airwaves daily, none commands more attention than the periodic reports of presidential popularity (Orren, 1978:35)."[1] Anecdotes of President Lyndon Johnson bandying about poll reports of his popularity are well known, and it is widely believed that the behavior of politicians is influenced by the behavior of these popularity measurements. As Brody and Page (1975:136-137) observe,

> . We can be confident that Presidents themselves are avid readers of the polls and view changes in the level of their popularity as feedback on the popularity of their actions. Lyndon Johnson was fond of quoting poll results to newsmen—at least in the early happy days of his administration. Both Johnson and Nixon earned reputations of eagerness to adapt their policies to the currents of public opinion. . . .
>
> Presidents, in short, have incentives to maximize their popularity; the polls provide information on how they are doing. Rises and falls in the polls after particular events may indicate what helps and what hurts and what corrective

actions should be taken. Popularity polls may therefore constitute part of a feedback system in which Presidents adjust their actions to public reactions.

More recently, commentators on the Carter presidency suggested that these measurements played an important role in some of that administration's more dramatic actions (see, for example, the account by Elizabeth Drew [1979] of President Carter's diagnosis of a "national malaise"; see also Converse [1979]). Indeed, one pundit was moved to write:

> Week after week we are treated to a succession of polls, each outbidding the other, with Carter Dropping, Carter Stabilizing, Carter Making Comeback, Carter Approaching Richard Nixon Lows. . . .
>
> [The corporate executives, editors and reporters who sponsor polls] believe that the noise they hear in the conch shell their pollsters hold up to their ears is the true, genuine, and unadulterated *vox populi*. So, unfortunately, do the politicans, including those in the White House.
>
> This explains the unbelievable events of the last couple of weeks in which a president, reacting to a bunch of percentages, hired and fired, made speeches, and did all manner of desperate things "to save his presidency," to use a phrase current among reporters. [Von Hoffman, 1979/1980:573]

The frequent reporting of these measurements was evidenced in the exploratory study of newspaper clippings that was reported in Chapter 2 of Volume 1. Reviewing American periodicals published in July of 1980, 40 reports of presidential-popularity polls were found (because of undercoverage in the sample, this figure considerably understates the actual frequency of such reports; see Appendix C of Volume 1). A review of national broadcasts of the evening news on the ABC, CBS, and NBC networks indicates that in the nonpresidential election years between 1973 and 1979, a monthly average of two to three nationally televised news stories were devoted to reports of such presidential-popularity readings.[2]

Presidential-popularity measurements, originated by the Gallup organization, extend back to at least 1941 (see Mueller, 1973:198). In later decades, four other survey organizations began publishing independent readings of presidential popularity. Louis Harris and Associates began reporting such measurements during the 1960s, and in the 1970s the Roper organization, the CBS News/*New York Times (NYT)* and NBC News/Associated Press (AP) polls initiated similar measurement programs.

The existence of a large body of measurements of the same subjective phenomenon made independently by five survey organizations presents a unique opportunity for assessing the reliability of subjective survey measurements. As noted in Chapter 5 of Volume 1, independent (and contemporaneous) replications of subjective survey measurements are not frequently made. Our search in other areas yielded only 120 instances since

1940 in which two survey organizations contemporaneously asked the same subjective question. Presidential popularity thus represents an unusual case, since we have 513 measurements of the same phenomena made independently by five survey organizations between 1963 and 1980. (As an indication of the scope of this enterprise, we would point out that more than 700,000 survey interviews were conducted to produce these measurements.)

Despite the political importance of presidential-popularity measurements, we are not aware of any previous attempt to systematically assess their interseries reliability. In the following pages we attempt that assessment, first by examining the extent to which the various series agree among themselves in their measurements of the level and variations across time in presidential popularity. Subsequently, we provide a separate evaluation of the extent to which the popularity data as a whole are responsive to factors that one might expect to affect the popularity of presidents. The latter exercise is not new in concept (see, for example, Mueller, 1973; Kernell, 1978; Monroe, 1978; Kernell and Hibbs, 1981), but our approach is somewhat novel in execution. In particular, by incorporating independent measurements made by five survey organizations, we are able to assess the relative sensitivity (and error) of each organization's measurements within the context of a substantive model of the dynamics of presidential popularity. The latter exercise provides us with a second and substantively grounded assessment of the relative accuracy of the different measurements of presidential popularity.

Description of the Data

Publicly available measurements of presidential popularity were collected from the following:[3]

1. Gallup: 313 measurements beginning in 1963
2. Harris: 116 measurements beginning in 1963
3. CBS/*NYT*: 23 measurements beginning in 1977
4. NBC/AP: 34 measurements beginning in 1977
5. Roper: 27 measurements beginning in 1973

Survey measurements made by these organizations used three basic question forms. The form used by both Gallup and CBS/*NYT* asked:

> Do you approve or disapprove of the way [name] is handling his job as president?

Louis Harris and Associates and NBC/AP used roughly similar question forms, which asked respondents to rate the job the president was doing. In the Harris version respondents were asked:

> How would you rate the job President [name] is doing as president—excellent, pretty good, only fair, or poor?

and NBC/AP asked

> What kind of job do you think [name] is doing as president—do you think he is doing an excellent job, a good job, only a fair job, or do you think he is doing a poor job?

Finally, the Roper organization used a third form of the question, which asked respondents:

> How do you feel about President [name] at the present time, would you describe yourself as a strong [name] supporter, a moderate [name] supporter, a moderate critic of [name], or a strong critic of [name]?

While the wording of each organization's question was meant to be invariant over time, our review of questionnaires filed at the Harris archive indicated that minor and apparently inadvertent alterations sometimes occurred. For example, the Harris item sometimes dropped the "as President" tag from the question (see Martin, McDuffee, and Presser, 1981:148-169).

While there is variation between survey organizations in the question posed (and minor variation within organizations), these presidential-popularity series are commonly thought to tap the same general phenomenon: public approval of the president's performance. It is thus reasonable to ask about the degree to which these five series produce comparable readings.

The dependent variable we have focused upon in our analyses is the (adjusted) proportion of respondents who approve of the president's performance. (In our contingency tables we use a dichotomy of approve versus disapprove, excluding all cases in which the respondent replied "don't know" or did not answer the question.) For each survey, the (adjusted) proportion approving (A) is:

$$A = \frac{\text{(Number approving)}}{\text{(Number approving + Number disapproving)}}$$

For the Harris and NBC/AP questions, the categories "excellent" and "(pretty) good" were called approval, and "fair" and "poor" were treated as disapproval. For the Roper question, "strong" and "moderate supporter"

were called approval and "strong" and "moderate critic" were treated as disapproval.

Our adjusted approval measure excludes "don't know" and "no answer" responses, because Sigelman (1981) has shown that the placement of presidential approval questions within the questionnaire (at least for Gallup's surveys) can alter the proportions of "don't know" (DK) and "no answer" (NA) responses. Sigelman reports, however, that question placement did not alter the ratio of approvals to disapprovals (when DK and NA responses were excluded). Since the placement of the popularity question within survey questionnaires has varied across time and organizations, our use of this adjusted rating should diminish the impact of any such artifacts.

Reliability Analyses

Initially we attempted to assess the extent to which the five time-series agreed with each other. This agreement might be termed the interseries reliability of the measurements. Agreement between the time-series produced by the five survey organizations can be understood in at least two senses. We can ask:

1. To what extent do organizations asking the same (or seemingly equivalent) questions obtain the same distribution of responses?

and we can also ask

2. To what extent do the various series of measurements (whether asking the same or only similar questions) obtain similar readings of the *changes over time* in public opinion? For example, when Organization 1's measurements of presidential popularity rise, do Organization 2's measurements rise in a similar manner?

Both of these questions are treated in this section.

Since presidential popularity changes with time, it would be of little use to compare measurements made in the spring by Organization 1 to those made in the summer by Organization 2. Of course, it is never possible to provide precisely overlapping measurement periods, but any reasonable analysis must come to grips with the need to ensure that variations in the timing of measurements are not overly large (or else one may risk mistaking the sensitivity of a series to the flow of events in time for simple measurement unreliability). In this reliability analysis, the criterion for temporal equivalence used a period of one-twentieth of a year in length (that is, about two-and-a-half weeks). We assume two measurements made during the same one-twentieth of a year period to be made at the same time. This assumption, of course, distorts reality to some extent. Dramatic events may

sometimes alter the public's regard for the performance of the president in a single day; the impact of such events will not be well represented by our procedures. It is nonetheless useful to remember in this regard that surveys themselves are not instantaneous. Telephone surveys are commonly conducted over a period of several days, and personal interview surveys often have field periods of more than a week. Hence, the surveys we discuss do not provide a fine-grained picture of public opinion; they too will mask the variations caused by such dramatic events. (This occurs because, for example, measurements made in interviews on the first of April will be included with those made in interviews conducted on the sixth of April.)[4]

DATA DESCRIPTION

We constructed a special dataset that included all instances in which two or more organizations made measurements at (approximately) the same time. We found 100 time periods during which two (or more) organizations asked the national population to evaluate their approval of the current president. We excluded all time periods in which only a single organization reported a measurement and all surveys of other than the national adult population (for example, samples of registered voters).

As noted, the 100 time periods were one-twentieth of a year in length.[5] The date of a survey was taken to be the midpoint of the sampling period. If two surveys were made by a single organization during the same time period, the results of the two surveys were aggregated.

Approval ratings were constructed from tabulation summaries provided by the Roper Public Opinion Research Center, the Louis Harris Data Center, and Louis Harris and Associates, Inc. Frequency counts of the numbers "approving" and "disapproving" were reconstructed using the percentage distributions and sample sizes provided by these archives. (Since these percentage distributions were reported in only two digits, a small amount of imprecision due to rounding errors was introduced into our reconstructed frequency counts.)

The reconstructed frequency counts are the basic data for the following reliability analyses.

COMPARING MEASUREMENTS MADE USING SIMILAR QUESTION WORDINGS

Identical Wording. The least-complicated reliability analysis involves measurements that were made using identically worded questions. The CBS/*NYT* survey has since its inception in 1977 used the Gallup wording of the presidential-popularity question. Thus, the comparison of these two series is not vulnerable to any (simple or interactive) effects attributable to differences in question wording.

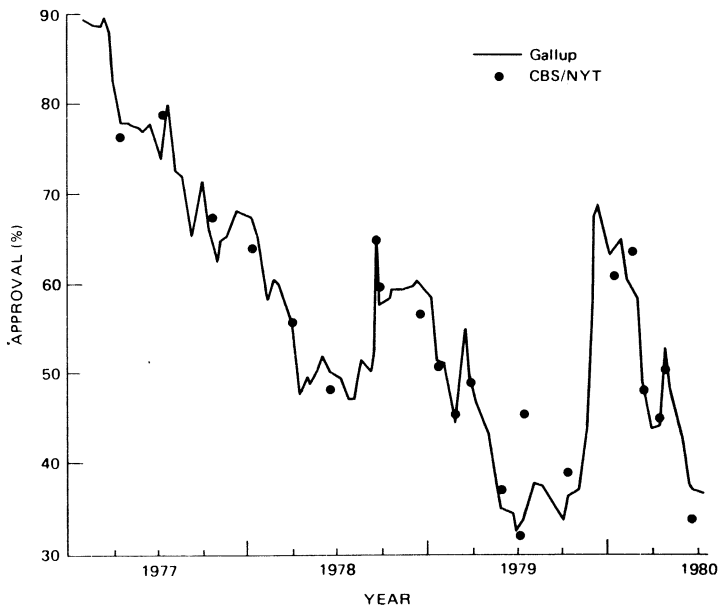
Accuracy of Presidential Ratings, 1963–80

In Figure 16.1 we trace the path of the Gallup (adjusted) approval rating during the years 1977–80 and insert points representing the 23 survey measurements made by the CBS/*NYT* poll. It quickly will be seen that the agreement between these two series is remarkably close. The product-moment correlation between the measurements made by the two organizations was 0.975 (for the 19 time periods in which both organizations made measurements). This correlation is comfortingly high, but it should be remembered that it reflects two factors: (1) the relatively large fluctuations evidenced by the underlying phenomenon, and (2) the relatively low level of disagreement between the series. Thus, although there is a very substantial agreement between the series about the large variations that occur in presidential popularity, it is still possible that at any given time the two series may give significantly discrepant readings of the level or the variation across time in the popularity of the president.

To study this possibility in greater depth, we fit five models to the 3-way table of frequency counts for the Gallup and CBS/*NYT* popularity measure-

FIGURE 16.1

Presidential Approval Measurements Made by Gallup (line) and CBS/NYT (dots), 1977-80.



NOTE: The time period of this plot is restricted to years in which CBS/*NYT* made survey measurements.

ments (that is, 19 time periods [T] by 2 survey houses [H] by 2 response categories [A: approve or disapprove]). These models were:

- Model 1: *No Change and Consistent Indicators.* Model 1 posits that the two organizations obtained identical estimates of a phenomenon that did not change over time. (Model constrained to fit {*A*} and {*HT*} marginals.)

- Model 2: *No Change and Constant Bias in Measurements.* Model 2 posits that the two organizations differed by a fixed amount between themselves in measuring a phenomenon that did not change over time. (Model constrained to fit {*AH*} and {*HT*} marginals.)

- Model 3: *Reliable Indicators of Change.* Model 3 posits that the two organizations obtained identical estimates of a phenomenon that changed over time. (Model constrained to fit {*AT*} and {*HT*} marginals.)

- Model 4: *Constant Bias in Measurements of a Changing Phenomenon.* Model 4 posits that the two organizations tracked the same variations across time, but their measurements differed by a constant amount at each time point. (Model constrained to fit all 2-way marginals, that is, {*AT*}, {*AH*}, {*HT*}).

- Model 5: *Variable Bias in Measurement of Changing Phenomenon.* Model 5 posits a 3-way interaction, that is, that the two organizations measured a phenomenon that changed with time, and that the bias in each organization's measurements also varied over time. (Entire 3-way {*AHT*} distribution fit.)

Table 16.1 presents the fit of each of these models to the data from the Gallup and CBS/*NYT* surveys. Since the wording of the questions used by these two organizations was identical, one would hope to fit Model 3 to these data. However, neither Model 3 nor the simpler models provide an adequate fit to the data collected by these two survey organizations. Moreover, somewhat different variations across time are observed in the two series. Hence, Model 4 does not provide an adequate fit ($L^2 = 52.6$, d.f. = 18, $p < .0001$). We are thus forced to accept a model that postulates a modest confounding of variations over time with those induced by survey-specific measurement biases.

We do note, however, that a model incorporating a parameter representing a *constant* house effect provides an insignificant improvement in the fit to the observed data (Model 1 minus Model 2: $L^2 = 3.5$, d.f. = 1, $p =$

TABLE 16.1

Test of Fit of Alternative Models for Behavior of Measurements of Presidential Popularity Made Using Identical Question Wordings: Gallup Versus CBS/NYT

Model for Behavior of Series	Mean Absolute[a] Discrepancy from Model	Test of Fit[b] L^2	Test of Fit[b] d.f.
1. No temporal change and no difference between organizations: {*A*}, {*HT*}	10.10	4,026.5	37
2. No temporal change, only difference between organizations: {*AH*}, {*HT*}	10.07	4,023.0	36
3. No difference between organizations, only temporal change: {*AT*}, {*HT*}	1.23	58.7	19
4. Temporal change and *constant* difference between organizations: {*AT*}, {*AH*}, {*HT*}	1.15	52.6	18
5. 3-way interaction (differences between organizations vary with time): {*AHT*}	0.0	0.0	0
Total N: 61,415			

NOTES: Gallup and the CBS/*NYT* surveys use identically worded questions. Measurements have been aggregated into approximately biweekly time periods (i.e., we have added together frequency counts for all measurements made in the same time period, each year was divided into 20 time periods).

A = Approval, *H* = Survey House, *T* = Time

L^2 = Likelihood-ratio chi-square

[a]Mean absolute discrepancy between observed frequency count and those expected under the model. This statistic (sometimes called the index of inconsistency) indicates the percentage of observations that would have to change (from approve to disapprove, or vice versa) in order to bring them into complete agreement with the values expected under the model.

[b]Except for Model 5, all tests of fit have $p < .0001$

.061). Moreover, we note that the relative improvement in fit that is obtained by fitting the 18 parameters representing all possible 3-way interactions of Approval by Time by Organization {*AHT*} is relatively small ($L^2 = 52.6$) given the enormous size of our sample. Finally, we note that the discrepancies (between the actual measurements made by the two survey organizations and those measurements expected under Model 4) are sufficiently small that only 1.15 percent of the measurements would have to change (from approval to disapproval, or vice versa) in order to bring the observed and expected measurements into complete agreement.

To provide a further understanding of the magnitude of the effects in-

TABLE 16.2

Estimates of Lambda Parameters Representing Effects of Time, House, and Time-by-House Interactions on Measurements of Presidential Popularity by Gallup and CBS/NYT

Lambda Parameters for		Estimate (and SE)		Standardized Value
House Effect (λ^{AH})		−.008	(.005)	−1.76
Time Effects (λ^{AT})				
1:	1977–06	.610	(.030)	20.32
2:	−11	.552	(.033)	16.70
3:	−17	.283	(.018)	16.03
4:	1978–01	.288	(.020)	14.08
5:	−06	.038	(.017)	2.18
6:	−10	−.050	(.020)	−2.55
7:	−15	.171	(.014)	12.01
8:	1979–02	−.019	(.017)	−1.11
9:	−04	−.133	(.019)	−7.02
10:	−05	−.025	(.018)	−1.41
11:	−09	−.333	(.021)	−16.14
12:	−11	−.339	(.016)	−21.05
13:	−16	−.304	(.018)	−17.21
14:	−17	−.295	(.021)	−14.33
15:	1980–01	.208	(.020)	10.55
16:	−05	−.122	(.019)	−6.31
17:	−06	−.149	(.019)	−7.99
18:	−07	−.042	(.020)	−2.14
19:	−10	−.339	(.017)	−19.66
Time-by-House Interactions (λ^{AHT})				
1:	1977–06 by Gallup	.068	(.030)	2.27
2:	−11 by Gallup	−.056	(.033)	−1.70
3:	−17 by Gallup	−.034	(.018)	−1.94
4:	1978–01 by Gallup	.046	(.020)	2.22
5:	−06 by Gallup	−.031	(.017)	−1.79
6:	−10 by Gallup	.028	(.020)	1.40
7:	−15 by Gallup	−.022	(.014)	−1.52
8:	1979–02 by Gallup	.008	(.017)	0.48
9:	−04 by Gallup	.004	(.019)	0.22
10:	−05 by Gallup	.039	(.018)	2.16
11:	−09 by Gallup	−.019	(.021)	−0.92
12:	−11 by Gallup	−.043	(.016)	−2.67
13:	−16 by Gallup	−.036	(.018)	−2.01
14:	−17 by Gallup	.001	(.021)	0.05
15:	1980–01 by Gallup	.034	(.020)	1.73
16:	−05 by Gallup	−.035	(.019)	−1.80
17:	−06 by Gallup	.003	(.019)	0.14
18:	−07 by Gallup	−.003	(.020)	−0.15
19:	−10 by Gallup	.048	(.017)	2.78

NOTES: Estimates derived from fitting of saturated model (No. 5). House effect shown is for Gallup. Time effects and House-by-Time interactions are differences between effect for indicated level and average effect for all levels. Time periods were one-twentieth of a year in length, so, for example, 1979–01 indicates the first twentieth of 1979 (approximately the first 18 days).

volved in producing these results, Table 16.2 presents estimates for the lambda parameters representing the effects of Time, House, and Time-by-House interactions on Approval. Lambda parameters were estimated using the saturated log-linear model (No. 5) for the 3-way table $A \times H \times T$. Following Goodman (1970:Sec. 3.1) we use P_{ijk} to denote the probability (in a population table) that an observation will fall in cell (i,j,k) ($i = 1, 2$ [approve, disapprove]; $j = 1, 2$ [Gallup, CBS/*NYT*]; $k = 1, 19$ [for time periods]), and $v_{ijk} = \log P_{ijk}$ (where log refers to the natural logarithm). The v_{ijk} are decomposed as follows:

$$v_{ijk} = \mu + \lambda_i^A + \lambda_j^H + \lambda_k^T + \lambda_{ij}^{AH} + \lambda_{ik}^{AT} + \lambda_{jk}^{HT} + \lambda_{ijk}^{AHT}$$

The lambdas represent the possible effects of variables A, H, T on v_{ijk}. For sample data, maximum-likelihood estimates of the lambdas can be obtained using procedures described by Goodman (1970). Viewing A as the dependent variable, the parameters that are of particular interest are the λ^{AH} and λ^{AT} parameters, which represent the 2-way interactions of Approval by House and Approval by Time, and the λ^{AHT} parameters, representing the 3-way interaction. The first two represent the "true" change that is occurring over time (λ^{AT}) and any constant difference that exists between the survey houses (λ^{AH}). In the present analysis, the last set of parameters (λ^{AHT}) represents effects that might be thought of as noise, that is, deviations induced by sampling variation *plus* other factors that have not been specified in our analysis, such as differences in sample design and execution, in the precise timing of surveys, in measurement procedures, and so forth.[6]

It can be readily seen from the estimates presented in Table 16.2 that

1. Notable House-by-Time interaction effects (estimated magnitudes exceeding twice their standard errors) occurred in 6 of the 19 time periods, and hence the inadequate fit of Model 4 shown in Table 16.1 is not due to a rare "odd" measurement.
2. However, the Time-by-House interaction effects, while significant, are of much smaller magnitude than the effects induced by Time.

· *Almost Identical Questions*. Since its inception in 1977, the NBC/AP poll has used a question that parallels the question used by the Harris organization. Both ask respondents to rate the job the incumbent is doing as president, and both provide four categories for response. The categories vary in one case:

Harris categories:	excellent	pretty good	only fair	poor
NBC/AP categories:	excellent	good	only fair	poor

We note too that the Harris question refers to the incumbent as "President"

(for example, President Carter), while the NBC/AP question refers to him by name only (for example, Jimmy Carter). Such differences in the wording of the two questions might lead one to expect minor variations in the response distributions obtained by Harris and NBC/AP. Notwithstanding these differences, one certainly would expect them to track parallel changes across time.

There were 17 biweekly time periods in which both Harris and NBC/AP made measurements. The product-moment correlation between the approval ratings obtained by the two organizations in those 17 time periods was 0.955.

Table 16.3 fits the five models described previously to the 3-way table of Approval *by* Organization *by* Time. It can be seen from these results that there is a substantial difference between the two organizations in their results—suggestive of a wording effect (or other house effect). Inclusion of a model parameter to allow a constant difference between organizations (that is, constraining the model to fit the {*AH*} marginals) produces a substantial improvement in model fit (Model 4 minus Model 3: $L^2 = 164.7$, d.f. = 1, $p < .0001$).

As with the CBS/*NYT* and Gallup comparison, a model incorporating both true change and survey-specific measurement biases is required, since Model 4 provides an inadequate fit to these data ($L^2 = 96.4$, d.f. = 16, $p < .0001$). We note also that the fit of Model 4 (true change plus constant bias) is poorer for these two series than for the Gallup-CBS/*NYT* comparison ($L^2 = 96.4$, and mean absolute discrepancy of 1.67 versus 52.6 and 1.15, respectively). This occurs despite the fact that the Gallup-CBS/*NYT* comparison included more time periods (that is, d.f. for the test of fit is 18 versus 16 for Harris-NBC/AP).

Table 16.4 presents estimates of the lambda coefficients representing the effects of House, Time, and Time-by-House interactions on Approval. It will be seen, as with the preceding comparisons, that there are many notable interaction effects; these effects are significant, but considerably smaller in magnitude than the effects induced by Time. Here, however, unlike the preceding comparison, we do find a notable constant effect of House on Approval (the lambda parameter for this effect has an estimated magnitude of 0.063, $SE = 0.005$). This suggests that in any naive comparison of two measurements made in different time periods by these organizations, the magnitude of any temporal change may be masked (or accentuated) by a noteworthy House effect (or wording effect)—as well as by significant House-by-Time interactions.

In considering this result (and those in Table 16.2), we might conceive of the "true" time-induced changes in presidential popularity as the "signal" that analysts are attempting to monitor. Survey measurements reproduce

TABLE 16.3

Test of Fit of Alternative Models for Behavior of Measurements of Presidential Popularity Made Using Similar Question Wordings: Harris Versus NBC/AP

Model for Behavior of Series	Mean Absolute[a] Discrepancy from Model	Test of Fit[b] L^2	Test of Fit[b] d.f.
1. No temporal change and no difference between organizations: {*A*}, {*IIT*}	10.10	3,604.1	33
2. No temporal change, only difference between organizations: {*AII*}, {*IIT*}	9.99	3,533.6	32
3. No difference between organizations, only temporal change: {*AT*}, {*IIT*}	2.78	261.1	17
4. Temporal change and *constant* difference between organizations: {*AT*}, {*AII*}, {*IIT*}	1.67	96.4	16
5. 3-way interaction (differences between organizations vary with time): {*AIIT*}	0.0	0.0	0
Total N: 52,481			

NOTES: The Harris Survey and the NBC/AP surveys used very similar question wordings. See text for exact wordings. Measurements have been aggregated into approximately biweekly time periods (i.e., we have added together frequency counts for all measurements made in the same time period; each year was divided into 20 time periods).

A = Approval; *H* = Survey House; *T* = Time

L^2 = Likelihood ratio chi-square.

[a] Mean absolute discrepancy between observed frequency count and those expected under the model. This statistic (sometimes called the index of dissimilarity) indicates the percentage of observations that would have to change (from approve to disapprove, or vice versa) in order to bring them into complete agreement with the values expected under the model.

[b] Except for Model 5, all tests of fit have $p < .0001$.

this signal with less than perfect fidelity. (Sampling fluctuations, of course, provide one component of "noise," but it is a component for which we have a well-developed theoretical framework to use in deciding when measurement variations are large enough to be attributed to variations in the signal.) Other (largely unknown) measurement factors also introduce a component of noise. Tables 16.2 and 16.4 present estimates of the magnitude of both the true variations that occurred in the signal (that is, the Time "effects") and estimates of the noise introduced by constant and time-variant

TABLE 16.4

Estimates of Lambda Parameters Representing Effects of Time, House, and Time-by-House Interactions on Measurements of Presidential Popularity by Harris and NBC/AP

Lambda Parameters for		Estimate (and *SE*)		Standardized Value
House Effect (λ^{AH})		.063	(.005)	13.28
Time Effects (λ^{AT})				
1:	1977 – 03	.769	(.025)	30.61
2:	– 10	.392	(.018)	21.75
3:	– 13	.272	(.018)	15.15
4:	– 16	.127	(.018)	7.19
5:	– 18	.075	(.018)	4.20
6:	1978 – 01	.070	(.019)	3.73
7:	– 03	– .078	(.018)	– 4.27
8:	– 05	– .064	(.018)	– 3.54
9:	– 07	– .220	(.020)	– 10.97
10:	– 10	– .214	(.017)	– 12.74
11:	– 15	.015	(.019)	0.79
12:	– 20	.001	(.018)	0.04
13:	1979 – 03	– .206	(.020)	– 10.36
14:	– 05	– .244	(.019)	– 12.92
15:	– 12	– .380	(.018)	– 21.45
16:	– 14	– .422	(.021)	– 20.35
17:	– 20	.107	(.019)	5.71
Time-by-House Interactions (λ^{AHT})				
1:	1977 – 03 by Harris	.105	(.025)	4.19
2:	– 10 by Harris	– .001	(.018)	– 0.03
3:	– 13 by Harris	– .088	(.018)	– 4.92
4:	– 16 by Harris	– .047	(.018)	– 2.67
5:	– 18 by Harris	– .037	(.018)	– 2.08
6:	1978 – 01 by Harris	.010	(.019)	0.54
7:	– 03 by Harris	.004	(.018)	0.20
8:	– 05 by Harris	.041	(.018)	2.27
9:	– 07 by Harris	– .013	(.020)	– 0.63
10:	– 10 by Harris	.039	(.017)	2.30
11:	– 15 by Harris	– .058	(.019)	– 3.08
12:	– 20 by Harris	.048	(.018)	2.69
13:	1979 – 03 by Harris	.027	(.020)	1.33
14:	– 05 by Harris	– .043	(.019)	– 2.29
15:	– 12 by Harris	– .014	(.018)	– 0.81
16:	– 14 by Harris	.065	(.021)	3.12
17:	– 20 by Harris	– .037	(.019)	– 1.97

NOTES: Estimates derived from fitting of saturated model (No. 5). House effect shown is for Harris. Time effects and House by Time interactions are differences between effect for indicated level and average effect for all levels. Time periods were one-twentieth of a year in length, so, for example, 1979 – 01 indicates the first twentieth of 1979 (approximately the first 18 days).

artifacts in our measuring instruments (that is, the House effect and the Time-by-House interaction effects).

Both Table 16.4 and Table 16.2 indicate that the signal-to-noise ratio in our measuring systems is quite favorable: the signal the survey organizations are monitoring shows fluctuations that are considerably larger, in general, than the noise introduced by our measuring instruments. Nonetheless, since there are detectable noises that distort our readings, it follows that comparisons of measurements that use (only) sampling error as a criterion for inferring changes in the signal must produce more errors of inference than would otherwise be expected.

Comparing Measurements Made Using Different Questions. Up to this point, our analysis has focused on differences between pairs of organizations that used (virtually) identical question wordings. This strategy ignores the rest of the presidential-popularity measurements and thereby makes it impossible to answer many questions of interest to us, for example, whether any one organization's measurements are widely discrepant from those made by other organizations. In the following section we treat these questions by analyzing the entire set of measurements for the 100 time periods.

As a first step in our analysis we computed the product-moment correlations[7] between the five measurement series. These results are presented in Table 16.5. It will be noted from this analysis that while the general level of the correlations is high, there is some variation between organizations. Measurements made by the Harris organization have the lowest mean correlation, although this correlation is still reasonably large (+0.879).

As noted previously, these correlations can obscure substantial discrepancies between the series, and so we undertook further analyses. To assess the reliability of these data, we fit models of the sort previously described

TABLE 16.5

Product Moment Correlations for Five Presidential-Popularity Time Series

	Gallup	Harris	CBS	NBC	Roper	Mean
Gallup	—	.928 (61)	.975 (19)	.939 (26)	.961 (22)	.951
Harris		—	.737 (7)	.955 (17)	.897 (11)	.879
CBS			—	.968 (10)	.986 (6)	.917
NBC				—	.937 (7)	.950
Roper					—	.945

NOTES: Approval rates obtained by each organization were aggregated into approximately biweekly time periods (1/20th of a year). Entries in parentheses show the number of time periods for which there are estimates by both organizations. (Coefficients derived from a weighted analysis, in which weights were proportional to survey sample sizes, produced virtually identical results; see note 7.)

to the full 3-way table representing the responses (that is, approve versus disapprove) obtained by the five survey organizations for all 100 time periods. This table, of course, is not complete, since all organizations did not make measurements in every time period. Indeed, of the 1,000 cells in the 3-way table, only 476 contain nonzero entries. Most of the missing data are concentrated in the period 1963–76; during that period only two organizations, Gallup and Harris, made regular measurements. Using the procedures of Goodman (1968; 1978:Chapters 4 and 5) we fit the hierarchy of models shown in Table 16.6 to the full 3-way table. Zeros were fit in all instances where observations were missing from the table.[8]

Table 16.6 provides strong evidence of the effects of time and organizational factors upon measurements of presidential popularity. Since question wording differs between organizations, we do expect to find differences in the levels of approval obtained by the different organizations. We would, nonetheless, hope to observe parallel changes across time in the data produced by each organization. Model 4, which fits the Approval-by-House $\{AH\}$ marginals (in addition to the $\{AT\}$ and $\{HT\}$ marginals), postulates the appropriate set of constraints. We observe two things about the fit of Model 4 to these data:

1. Tests of fit for Model 4 provide evidence of the constant effects of wording (and/or other organizational factors) on response. The existence of these effects can be tested by contrasting the fit of Model 3, which does not constrain the $\{AH\}$ marginals and Model 4, which does. The effects are substantial and statistically reliable (Model 3 minus Model 4: $L^2 = 6{,}478$, d.f. $= 4$, $p < .0001$; mean absolute discrepancy: 5.57 versus 2.21).
2. And, we note that Model 4 does not provide an adequate fit to the observed data ($L^2 = 1{,}452$, d.f. $= 134$, $p < .0001$). Despite the lack of a fully acceptable fit, we do note that the mean absolute discrepancy[9] between Model 4 and the observed data declined to the point where only 2.21 percent of the sample would have to change in order for the observed frequency counts and the counts expected under the model to be in perfect agreement. [When wording (plus organization) effects are excluded from the model, the corresponding percentage rises to 5.57.]

Our failure to obtain a satisfactory fit of Model 4 to the pooled set of measurements naturally raises questions about the relative contributions of each organization's measurements to this lack of fit. One may wish to know whether the measurements made by one (or more) organizations are consistent outliers. To assess this, we partitioned the L^2 statistic for the fit of Model 4 into components contributed by each organization.[10] Table 16.7 displays the results of this analysis.

It will be seen from Table 16.7 that one organization (CBS/*NYT*) shows a

TABLE 16.6
Test of Fit of Alternative Models for Measurements of Presidential Popularity Made by Five Survey Organizations

Model for Behavior of Series	Mean Absolute[a] Discrepancy from Model	Test of Fit[b] L^2	d.f.
1. No differences across time or organizations: $\{A\}$, $\{HT\}$	13.21	40,655.4	237
2. Only differences between organizations: $\{AH\}$, $\{HT\}$	11.73	33,906.0	233
3. Only differences across time: $\{AT\}$, $\{HT\}$	5.57	7,930.4	138
4. Parallel changes across time (but *constant* differences between organizations): $\{AT\}$, $\{AH\}$, $\{HT\}$	2.21	1,452.3	134
5. 3-way interaction (differences in approval between organizations vary with time): $\{AHT\}$	0.0	0.0	0
Total N: 378,043			

NOTES: Measurements have been aggregated into approximately biweekly time periods (i.e., we have added together frequency counts for all measurements made in the same time period, each year was divided into 20 time periods).
A = Approval, H = Survey House; T = Time
L^2 = Likelihood ratio chi-square.

[a] Mean absolute discrepancy between observed frequency count and those expected under the model. This statistic (sometimes called the index of dissimilarity) indicates the percentage of observations that would have to change (from approve to dissapprove, or vice versa) in order to bring them into complete agreement with the values expected under the model.

[b] Except for Model 5, all tests of fit have $p < .0001$

deviation from model expectations that is about as large as would be expected on the basis of sampling fluctuations alone. The 21 CBS/*NYT* measurements contribute just $L^2 = 32.4$, and only 1.36 percent of the CBS/*NYT* observations would have to change to bring them into complete agreement with the expectations derived under Model 4. In contrast, the 68 measurements of the Harris organization are most discrepant (mean absolute discrepancy = 2.77 percent; contribution to $L^2 = 538.4$).

In interpreting the results shown in Table 16.7, it should be kept in mind that the expected value for an observation is a function of the central tendency of the entire body of observations at any given time point. When there are only two organizations making measurements in a time period, it is theoretically impossible to distinguish the instability of one set of measurements from the rectitude of the other. Since only two organizations

TABLE 16.7
Deviations of Five Survey Organizations' Measurements from Values Expected Under Model 4

Time Period and Organization	Mean Absolute[a] Discrepancy from Model	L^2	N	
			Measurements	Respondents
1966 – 80				
1. CBS	1.36	32.4	21	24,478
2. Roper	1.74	95.7	26	46,937
3. NBC	2.09	167.7	31	49,632
4. Gallup	2.15	618.1	92	156,910
5. Harris	2.77	538.4	68	100,086
All Organizations	2.21	1452.3	238	378,043
1977 – 80				
1. CBS	1.36	32.4	21	24,478
2. Roper	1.60	57.7	18	33,052
3. Gallup	1.77	215.2	51	96,565
4. NBC	2.09	167.7	31	49,632
5. Harris	2.53	194.8	32	48,166
All Organizations	1.92	667.8	153	251,893

NOTES: Discrepancies are differences between expected values under Model 4 (see Table 16.6) and observed values for each organization.

L^2 values show each organization's contribution to likelihood-ratio chi-square statistics for deviation of observations from expectations under Model 4 (see Table 16.6)

[a] Mean absolute discrepancy between observed measurement and that expected under Model 4. This statistic (sometimes called the index of dissimilarity) indicates the percentage of observations that would have to change (from approve to disapprove, or vice versa) in order to bring them into complete agreement with the values expected under the model.

(Gallup and Harris) were active during the period 1962–76, the inconsistency between these organizations affects both of them equally.

To provide more useful information on the relative performance of the five survey organizations, we separately analyzed the residuals for the 1977–80 period. During this period the pooled data from the other three organizations can be used to adjudicate disagreements between Harris and Gallup. (This approach is roughly equivalent to asking which organization best matches the measurements made by the other three.)

Using the 1977–80 results, we note that the Gallup organization improves its relative ranking, while Harris remains the most discrepant of the five organizations (mean absolute discrepancy from model = 2.53 percent). Indeed, we note that the Gallup and Harris organizations contribute almost equally to our failure to fit Model 4 during this time period (L^2 = 215.2 and 194.8), even though Gallup made 50 measurements (N = 96,565) compared to 30 for Harris (N = 48,166).

Overall, the average magnitude of the observed discrepancies in the entire set of measurements is 1.92 percentage points for the years 1977

through 1980. Three organizations (CBS/*NYT*, Roper, and Gallup) have average discrepancies that are smaller than this, ranging from 1.36 to 1.78 percentage points. The remaining organizations (NBC/AP and Harris) show larger discrepancies. In the case of the Harris organization, its average discrepancy is almost twice that of the organization with the least discrepant measurements, CBS/*NYT* (2.53 versus 1.26 percentage points).

Fitting a Substantive Model

The direct comparison of popularity measurements generates a precise test of survey-house variability. The only difficulty stems from the arbitrariness of the central tendency for each time point against which each house's measurement is compared. In theory, that comparison point is determined by the "true" popularity at the moment of the simultaneous measurements; in practice, it is determined endogenously, which is to say, by the measurements themselves. The test does not explicitly take into account the fact that presidential popularity ought to be generated by the political and economic environment at any time point, and that an exogenously determined "popularity" may be arrived at. Such a comparison point for the house measures can provide an additional standard for comparison. That is, an external-construct validity test may be added to the reliability test for cross-house comparisons.

The strategy used in this section introduces a substantive causal element by modeling presidential popularity as a dynamic function of the political environment. A cross-house difference can then be tested as a direct, but separable, effect. The idea is to first develop a full model of presidential popularity as a response to the conditions of the economy, external warfare, and ordinary political events, and then to compare instrument differences which cannot be attributed to variations in the contemporaneous environment. Finally, a further assessment of measurement reliability may be added to the direct comparison by analyzing the residual variation left after the substantive and (constant) house effects have been estimated.

CONSTRUCT VALIDITY

The substantive model of popularity used here is one which is about as simple as possible and which yet can capture the flavor of the short-term dynamics found in these data. For each variable in the model, popularity is taken to be a reflection of the current level of that variable and, with decreasing weight, its past levels. For example, support for the president is assumed to be (partially) due to today's inflation rate, to a lesser extent due to last month's inflation rate, and to an even lesser extent due to the previous month's, and so on. A simple way to write such a function is

$$P_t^{(i)} = bI_t + dP_{t-1}^{(i)}$$

where $P^{(i)}$ is the component of popularity attributable to inflation; I_t is the level of inflation in the tth time period. This equation represents a familiar partial-equilibration process in which the effect of any shock (in this case the level of inflation) will persist into the future as an exponential function of the passage of time (bd^t) as the population's judgment of the president slowly returns (equilibrates) to a level determined by the rest of the system. This first-order model represents a process in which the rate of reequilibration is a simple linear function ($1 - d$) of the extent to which the system is out of equilibrium at any moment. This model yields a response pattern that begins at its maximum and then gently, exponentially, approaches the equilibrium level as an asymptote.

By assuming that popularity is an additive function of the factors in the model, estimates for the scalar transforms (the b's) and the dynamic coefficients (the d's) may be obtained for each variable. The straightforward Koyck transform (Koyck, 1954; Theil, 1971) cannot be used for these estimation purposes because (1) the previous popularity level, P_{t-1}, is not uniformly available for each reading; (2) using it would introduce the effect of previous, unspecified disturbance terms into the dynamic coefficients; and (3) the simple model assumes equal equilibration rates for all forms of system shock. The first problem is a practical one and might otherwise be circumvented by collapsing the series into a monthly one (although this would limit the analysis to periods of frequent measurements). The second problem is a standard estimation bias and needs attention; the procedure here solves the problem only in part. More important, the third problem is that a simple Koyck transform depends on an assumption that is theoretically unsatisfactory. We might naturally expect the effect of, say, a presidential speech to be more transitory than that of, say, an election campaign, simply because the latter is likely to generate a complex attitude change, more resistant to subsequent reequilibration forces. This is not to say that the immediate impact will be greater (though it might be), but that its observable duration will be longer.

With this in mind we obtained the current results by directly estimating the model's dynamic parameters with a nonlinear procedure coupling a grid search and Marquardt's iterative gradient compromise. In practice this amounted to setting initial conditions for 1961 and dynamically generating the estimated popularity path for subsequent time points as the specified function of the previously estimated popularity and contemporaneous input.[11] The lay reader will note that the observed values of popularity (for example, P_{t-1}) never appear on the right-hand side of the estimation equation as they are replaced by values estimated from previous shocks to the system. A further advantage of this estimation strategy lies in the fact that

modeling presidential popularity on exogenous sources allows estimates to be obtained for all data points,[12] not just those for which another house made a coincident reading (as was the case in our previous analyses).

Substantively there is nothing remarkable about the underlying model: it is a linear combination of four sets of dynamic components plus a dummy variable for each presidential administration.[13] The economic variables, unemployment and inflation, were introduced additively[14] (experimentation with interactive components did not compel a more complex specification). The Vietnam War variable represents the number of troops (in tens of thousands) in Vietnam and thus provides an indicator of ordinary citizens' personal involvement with the war effort. This variable was set to zero after the inauguration of Nixon in the belief that Vietnam was not viewed as his war and because, empirically, these data did not show a relation between troop levels and Nixon's popularity. In addition, one other variable was employed. It represents the ephemeral glow that surrounds each new president as he takes office and citizens (appear to) give him the benefit of the doubt (followed by the subsequent gradual decline as doubts become justified). In practice this variable is simply a unit spike placed at inauguration day with its magnitude and decay rate left to be estimated from the data. Finally, in order to capture the short-term dynamics, two variables representing events of transitory and then of more long-lasting impacts were introduced.[15]

Table 16.8 displays the results from estimating this model,[16] assuming that the differences between each organization's measurements could be represented as a constant bias. The expected value at any time might vary by a constant effect if the the organization's survey procedures were constantly different with respect to wording, sampling frame, and so forth. In terms of wording, the effect might be due to the various probes and response categories given the respondent, or to an organization's interviewer training, that gave different instructions on the tone in which the question was to be read. In terms of the sampling frame, such a difference might obtain if the houses incorporated a constant bias in drawing their initial sample (say, picking more middle-class or better-educated respondents) or in their call-back procedures (which might overweight the sample with accessible respondents). The house coefficients in the estimation equation are dummy variables comparing each condition with Gallup sampling, wording, and interview procedures.

By far the most important result is that the CBS/*NYT* replication of the Gallup approval wording produces almost exactly the same estimates obtained by Gallup. This is to say that for any given reading of popularity, their *expected value* would be almost identical. (These estimates do not assess their relative precision; the matter of reliability requires analysis of variation, not expected value, as will be discussed below.) Here the data

TABLE 16.8
Presidential Popularity and Constant House Effects (Modeled as a Function of Political Climate, Administration, and Survey House)

Coefficients for		*B*	*SE(B)*	*D*	*SE(D)*
Intercept		72.53*	2.79		
House effects:	CBS	−0.05	1.25		
	Roper	0.52	1.05		
	Harris	−11.85*	0.67		
	NBC	−18.62*	1.02		
Substantive Model					
Presidential Administration:	Johnson	5.51*	2.56		
	Nixon	−9.70*	1.72		
	Ford	−6.56*	2.14		
	Carter	−17.26*	2.33		
Presidential glow:	Johnson	21.95*	3.82	.96	.32
	Nixon	37.09*	2.06	.96	.06
	Ford	53.17*	2.89	.85	.07
	Carter	53.77*	2.21	.97	.07
Political climate:	Short Events	1.09*	.05	.83	.07
	Long Events	1.06*	.09	.94	.11
	War	−0.12*	.01	.69	.07
	Inflation	−0.35*	.02	.84	1.40
	Unemployment	−0.65	.52	.00	—

NOTES: This model predicts all observations (except for voter samples), 1963–80. Variables were entered as shown, with the Constant House Effects introduced as dummy variables contrasted with Gallup as the baseline. Administration coefficients mark the president's equilibrium value that is not determined by the rest of the equation; they are contrasted with Kennedy's equilibrium popularity represented in the constant term. The fit may be indicated by: R-square = 0.91; F-ratio = 261.08; d.f. = (17/419), mean squared error = 24.29.
*Probability that parameter is equal to zero is less than .01

provide a good experiment on the comparability of survey methodologies in that the question wordings are identical and any difference in readings could be attributable to constant house-induced effects. The fact that two houses (CBS/*NYT* and Gallup) produce indistinguishable readings suggests that houses can eliminate biases in measurement, and this reinforces the contention that these measurements can be taken seriously. This conclusion must be taken all the more seriously because the similarity in readings may not be ascribed to each house's being aware of an experimental comparison, and thus paying particular care to its procedures; instead, the readings were produced by the houses' everyday practice.

More surprising, although more of a curiosity, is the inability to distin-

guish statistically the Roper results from those of Gallup and CBS/*NYT*. Both measures do emphasize a "pass the mark" frame of reference rather than a graded form of evaluation, and in doing so may produce similar readings. However, given the two very different sets of words used to elicit a response, we can only suggest that the correspondence be seen as a happy coincidence for the organizations and that it provides a puzzle for more elaborate questionnaire research.

We may not conclude that the idea of presidential approval is insensitive to measurement technique. This can easily be seen by inspecting the large differences in Table 16.8 for the Harris and NBC/AP measurements, which ask for a graded job rating, not support or approval. Clearly far fewer citizens are willing to judge the president's performance as excellent or (pretty) good than are willing to approve or support him, and the difference is statistically reliable. (The further distinction $[(-11.85) - (-18.62) = 6.77]$ between the very similar Harris and NBC/AP probes is troubling, but no firm conclusion is possible about its source, given the present data.)

The foregoing results are based on the assumption that the inconsistencies in measurement can be expressed as a constant bias. There is no reason to exclude a priori the possibility that the observed differences are due to time-variant survey procedures. For example, if an organization should have oversampled the highly educated or attentive public, then it may also oversample responsiveness to changes in the political climate. Alternatively, the different question formats may elicit a sentiment that is peculiarly sensitive to elements in the political and economic environment. Thus further examination is called for. In particular, the contrast between responses to the Gallup-CBS question and the Harris-NBC question may be due to a number of respondents who are willing to approve of a president whose job they rate "fair." (A presidential job rating of "only fair" is counted as not approving of his performance. See previous discussion.) If this were so, there would be a number of Gallup "approvers" whose support would be tentative in that they rate job performance as "only fair." These soft supporters might thus be more easily moved in and out by shifts in the political climate.

This notion of a proportion of Harris "only fair" raters also approving of the president has been studied. Sussman (1978) and Orren (1978) report a *Washington Post* survey in January 1978 directly comparing respondents' answers to both the Gallup and Harris questions. The *Post* found that 46 percent of those rating Carter's job as "only fair" also "approved" of his handling of the job. This same proportion may be estimated from our data across all the Gallup and Harris samples. The difference between the Harris and Gallup readings may be expressed as a proportion of those saying "only fair" to Harris who would "approve" in the Gallup format. Performing the proper analysis on the *Post* data, which is to say excluding the "don't

know" categories and working from the marginals, yields an estimate of 46.06 percent.[17] A similar procedure for all the Gallup-Harris direct comparisons for the 1969–80 period yields the estimates shown in Table 16.9, which are slightly smaller than the *Post*'s 1978 figure. Either the proportion was high for that reading or, more likely, the aggregate estimates suffer from some underspecification and thus are modestly depressed toward zero. Of some interest is the low estimate for the Watergate period, because it suggests that at that time only the president's hard-core supporters were willing to give Gallup an "approval" response. The overall pattern indicates that the difference in wording may have something to do with the different levels of popularity readings for the two questions, and it suggests that this difference may be due to a portion of presidential supporters whose support may be more susceptible to influence.

This possibility may be formally assessed by modeling the Gallup and Harris popularity measures with the coefficients free to pick up a unique sensitivity to the political-economic environment. The coefficients displayed in Table 16.10 are produced by a test of such a less restricted model of Gallup approval and Harris rating results for the Nixon, Ford, and Carter years.[18] Dummy variables (for Harris) were introduced to allow the substantive model to have distinct effects for each organization's measures of support. The first column presents the coefficients for the Gallup popularity readings. The second column of figures is the crucial set because it presents the magnitude of the *difference* in Gallup-Harris readings due to variable sensitivity to the model's substantive components.

Of those substantive coefficients, the only one that seems significantly different for the Harris readings is the one due to inflation. As expected, Harris's evaluative rating measure is less sensitive. Lukewarm supporters included by the Gallup question do seem more sensitive to fluctuations in

TABLE 16.9

Estimates of Proportion Rating President's Job "Only Fair" (Harris) Who "Approve" (Gallup)

Administration	Number of Comparisons	Proportion "Fair" Who "Approve"	Standard Error of Estimate
Nixon (before Watergate)[a]	13	.46	.08
Nixon (during Watergate)[b]	8	.05	.03
Ford	12	.38	.04
Carter	25	.37	.02

NOTES: These estimates derive from separate mappings of the Harris "only fair" sample proportion to the difference between Gallup and Harris approval ratings. Numbers are rough due to likely underspecification.
[a]I.e., prior to 15 March 1973.
[b]I.e., from 15 March 1973, to Nixon's resignation.

TABLE 16.10

Presidential Popularity and House/Substance Interaction Modeled on Political Climate, Administration, and Survey House (Gallup and Harris)

Coefficients for		Gallup: *B*	Harris Difference: *B*	*SE*: Difference
Intercept:		59.80*		
Presidential administration:	Nixon	—[a]	−16.69*	4.60
	Ford	2.06	−12.45*	3.38
	Carter	−5.31*	−13.34*	2.07
Presidential glow:	Nixon	43.49*	−14.10*	1.51
	Ford	63.30*	−11.46*	4.46
	Carter	52.87*	2.84	3.79
Political climate:	Short Event	1.12*	0.00	0.08
	Long Event	1.16*	0.03	0.15
	Inflation	−0.41*	0.18*	0.03
	Unemployment	0.37	1.30	1.08

NOTES: These estimates are for the Nixon, Ford, and Carter administrations only. Column 1 (Gallup) gives the baseline substantive model. Column 2 (Harris Difference) presents an estimate of the difference in the sensitivity of the Harris measure to the particular variable. The standard error of the estimate of the difference is presented in column 3. The analysis uses the same dynamics as those generated in the analysis reported in table 16.8. The fit may be indicated by: R-square = 0.97; F-ratio = 412.86; d.f. = (19/244); mean squared error = 8.01.
[a]The Nixon administration was the residual category for the comparison.
*Probability that parameter is equal to zero is less than .01.

the Consumer Price Index than do the relative stalwarts Harris picks up. Surprisingly, though, this differential sensitivity cannot be identified for the political event series or for unemployment (though the estimates of the latter are very imprecise). In any case, the differences here, though reliable, are very small.

Overall, however, the contrast between Gallup and Harris does not reflect unique sensitivities to the immediate environment, but instead is best represented as a simple function of the presidential equilibrium level and the glow for each administration. Gallup's question seems to capture consistently more approvers whose support is not affected by short-term changes in political climate and whose support erodes as the glow vanishes.

This form of comparison may be expanded to include all organizations, but only for the 1977–80 period. For this purpose, a separate model was "custom-fitted" to the popularity readings of each house (the collinearity in a full house/substance interaction model is overwhelming). Coefficient estimates, displayed in Table 16.11, are a little uncertain because Carter's readings moved so as to confound the contributions of the unemployment and glow variables. Nevertheless, the different questions' responsiveness

may be compared for the event and inflation variables. The same pattern is observed in that the Harris, and now the NBC/AP, questions seem less sensitive to inflation, but the statistical test of that difference is a little more murky. Their susceptibility to the impact of political events seems much the same as that for the other houses. In addition, it is evident that the Roper question (about presidential "supporters") is mildly less responsive to the event data. This is not surprising, given the loyalty implications of the response categories, but this difference is only on the edge of statistical significance.

This further work underscores the apparent distinction in Table 16.8 between Harris on the one hand and NBC/AP on the other. Estimated equilibrium levels for Carter seem noticeably different for those two houses, although the estimates are imprecise. A more explicit comparison in which the two organizations' readings for the Carter period are estimated together (thus constraining the model somewhat) yields an estimate of 6.49 for a constant difference[19] in their readings, with a standard error of 0.75. The probability that this difference is actually zero is considerably less than .01. Thus it is likely that there remains a distinction between the organizations' readings which cannot be attributed to their questions being differentially sensitive to the substantive determinants of popularity.[20]

TABLE 16.11

Presidential Popularity and House/Substance Interaction Custom Fitted to the Political Climate (Carter administration only—all houses)

	Coefficients for					
Organization	Intercept	Presidential Glow	Short Event	Long Event	Inflation	Unemployment
Gallup	53.73	55.15	1.09	1.06	−0.36	−0.26
	(3.24)	(2.20)	(0.05)	(0.13)	(0.03)	(0.54)
Harris	7.51	54.39	1.00	1.08	−0.10	4.17
	(19.03)	(12.59)	(0.15)	(0.26)	(0.07)	(5.41)
CBS/*NYT*	46.64	66.34	1.00	1.03	−0.22	−1.23
	(16.18)	(17.29)	(0.09)	(0.37)	(0.14)	(1.28)
NBC/AP	19.29	53.23	0.82	1.04	−0.16	0.94
	(7.36)	(4.94)	(0.12)	(0.24)	(0.08)	(1.49)
Roper	50.12	41.52	0.79	0.78	−0.34	1.96
	(13.11)	(9.11)	(0.09)	(0.18)	(0.06)	(3.91)

NOTES: These coefficients derive from a separate, custom-fitted model for each house for the Carter years. Intercept (equilibrium), glow, and unemployment coefficients are very imprecisely estimated due to their marked collinearity. More reliable are the estimates for the event series and for inflation. The standard errors of the coefficients' estimates are included in parentheses under each estimate. The model uses the same dynamics as that reported in table 16.8.
R^2 values were: .97 (Gallup), .95 (Harris), .93 (CBS/*NYT*), .95 (NBC/AP), and .98 (Roper).

A SECOND TEST OF RELIABILITY

The equations represented in Tables 16.8, 16.10, and 16.11 attempt to establish the fact that popularity measures do in fact covary with the external environment and that they do so in sensible and predictable ways. They do not assess how accurately these organizations measure popularity.

Another measure of the variability of these measurements, in addition to the direct test presented in our first set of reliability analyses, may be derived from analyzing the deviations of each organization's measurements from the score predicted by the substantive model. The utility of this effort depends on the predictions being meaningfully related to theoretical expectations; the foregoing analysis of construct validity seems to suggest that this is so.

A simple comparison of residual deviations is presented in Table 16.12, in which the predicted scores are manufactured by the constant-effects model of Table 16.8. The first column of numbers is the set of mean absolute values of the residuals associated with each survey house.[21] A first glance indicates that the most reliable organizations, that is, those whose average deviation was smallest, were CBS/*NYT*, Gallup, and Roper, while falling somewhat behind were NBC/AP and Harris. Comparing across houses, however, is a difficult problem, because the houses attempted to measure popularity during somewhat different periods. In particular, only Gallup and Harris generated frequent measures for any but the Carter years. Thus the second array of numbers displays the mean absolute deviations for each house from the constant-effects models reestimated for the 1977–80 period. The same classification of relative precision holds, though the scores look more accurate than before because of the model's superior prediction of the Georgian's popularity.[22] Differences within each clustering are not statistically significant, but the wide gap between Gallup, CBS, and Roper on the one hand and NBC and Harris on the other is large and statistically reliable. These results mimic the direct comparisons of the first section, and they add an extra weight to those findings because these are differences due not to an endogenously determined "true" value, but to one that has some substantive interpretability.

There is one potential source of error in this analysis, however. The constant-effects model does not take into account the minor question/substance interaction displayed in Table 16.10 for Harris and in Table 16.11 for Harris and NBC. Finding that both Harris and NBC fall toward the bottom suggests that the model may attribute to unreliability, error due to model underspecification. The predictive model is dominated (in terms of sample points) by the Gallup series, and thus the other house measures may be slighted by the fact that they may respond differently to environmental stimuli. Thus, an examination of the custom-fitted models represented in Table 16.11 is necessary in order to eliminate this possibility. Table 16.13

TABLE 16.12

Analysis of Residual Deviations from Constant-House-Effects Model: Popularity Modeled on Political Climate, Administration, and Survey House

Time Period	Organization	Mean Absolute Deviation	Standard Error of Mean	L^2	Number of Readings
1963–80	1. CBS	2.11	.45	87.0	20
	2. Gallup	2.33	.10	1905.4	270
	3. Roper	2.59	.44	311.4	27
	4. NBC	3.34	.47	458.1	31
	5. Harris	3.87	.34	1312.4	76
	All	2.69	.19	4074.3	424
1977–80	1. Roper	2.07	.41	108.4	18
	2. CBS	2.09	.45	86.5	20
	3. Gallup	2.38	.20	533.8	70
	4. NBC	2.66	.46	365.8	31
	5. Harris	3.10	.41	341.9	32
	All	2.53	.29	1436.5	171

NOTES: Entries in column 1 represent the absolute value of the residual deviations from the substantive predictive model. Models here represent the Constant-Effects model of the political climate, administration, and house shown in table 16.8. The difference between these tests and those of table 16.7 lies in the estimated polularity being derived from the substantive model. The second column presents the standard error of the mean estimates, and the third gives the likelihood-ratio chi-square. The last column shows the number of presidential-popularity readings taken by each house during the period. (Model parameters were separately estimated for the entire period, 1963–80, and for the years 1977–80.)

presents the residual analysis. Clearly, the same clustering of houses is manifest and the same statistical conclusion must be made. Whatever its source, both Harris and NBC consistently display significantly greater variation in their measurements.[23]

(Plots of these observed and expected presidential-popularity ratings for 1977 through 1980 are shown in Figures 16.2 and 16.3; in these figures the custom-fitted model was used to provide the predicted values for the measurements made by the five survey organizations. Figure 16.4 presents similar plots for the Gallup and Harris measurements during 1963 through 1976 using the constant-effects model.)

While the gap between the two reliability clusters may be identified, in absolute terms it is not large. After all, popularity ratings move over a range of about 60 points. However, in terms of a total-error model (see Volume 1, Chapter 4) the proportion of variability added on top of a theoretical sampling error is important. The expected deviation due to sampling runs from 0.88 to 1.05 for the organizations' Carter ratings, but the observed deviations range from 1.29 to 2.55.[24] And we note that Harris and NBC/AP add

TABLE 16.13

Analysis of Residual Deviations from Models Custom-Fitted for Each Survey House: Popularity Modeled on Political Climate (Carter administration only)

Time Period	Organization	Mean Absolute Deviation	Standard Error of Mean	L^2	Number of Readings
1977–80	1. Roper	1.29	.27	42.3	18
	2. CBS	1.80	.46	73.9	20
	3. Gallup	2.19	.18	456.9	70
	4. Harris	2.68	.34	246.0	32
	5. NBC	2.55	.41	265.0	31
	All	2.20	.25	1084.2	171

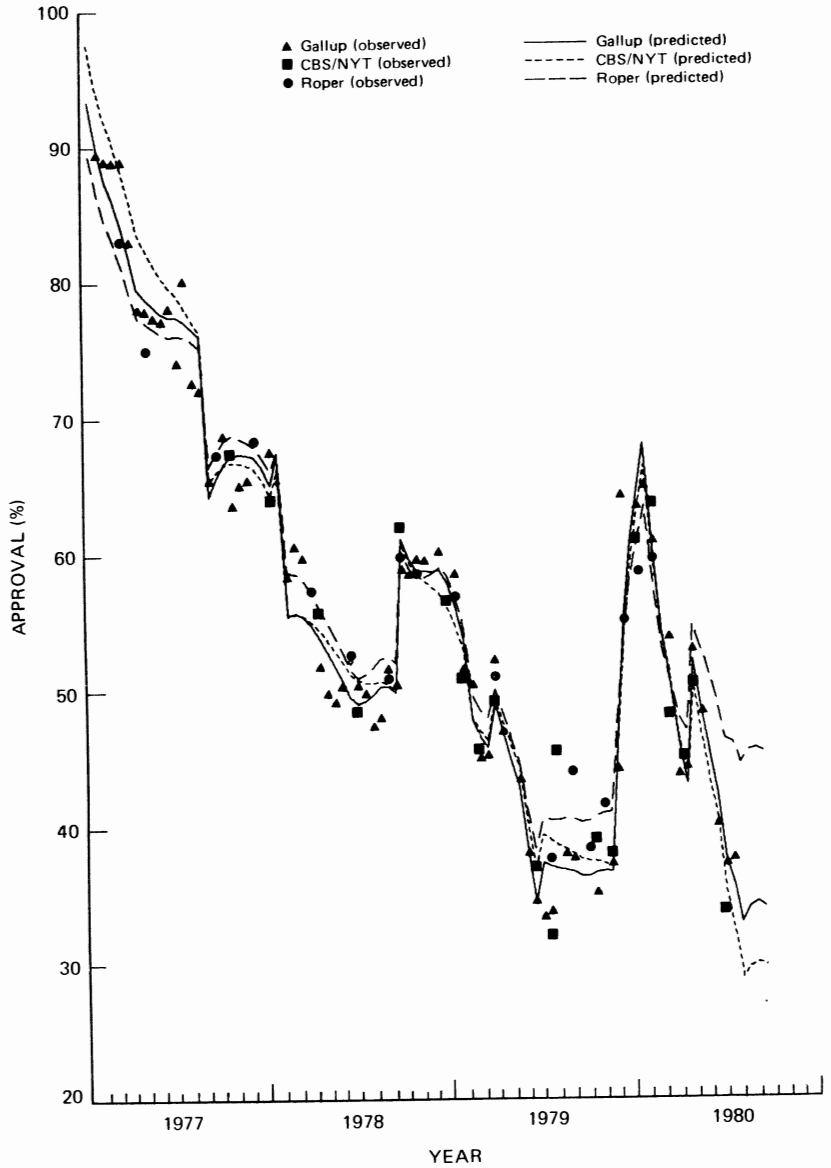
NOTES: Entries in column one are the absolute value of the residual deviations from the substantive predictive model. Here, a separate model was fitted for each house on the political climate (all for the Carter administration). The second column presents the standard error of the mean estimates, and the third the likelihood-ratio chi-square. The last column shows the number of popularity readings taken by each house during the period.

(very roughly) about 1.5 to 2 times the amount of measurement error that Gallup and CBS/*NYT* introduce.

The source of the latter difference and its implications for development are not clear. It could be that both NBC and Harris use survey practices that are noticeably less precise than those used by other organizations. This might be the result of less rigorous procedures for sampling, call-backs, and interviewer training. (The one available comparison in this regard [see Figure 3.1 in Volume 1] does indicate that the Harris samples substantially overrepresent the college-educated; no systematic evidence is available to us concerning the NBC/AP surveys.) However, finding that it is the houses with a particular question format that produce the greater variance suggests that the source may lie in the nature of their questions. Simply increasing the number of options presented to the respondent might be expected to increase test-retest reliability if there is any randomness in response. Furthermore, presentation of the "only fair" option may introduce an ambiguous response category that makes the division between, say, "pretty good" and "only fair" an easy line for uncertain respondents to slip over.[25] If the analyst were certain that respondents have a dichotomous attitude (approve or disapprove), the fine-grained response format may build in a measurement error. However, if respondents' attitudes genuinely fall along a continuum, the increased variability may be due to the respondents' uncertainty about how their feelings can be translated into simple approval or disapproval. If this were the case, the Gallup-CBS approval-disapproval forced choice may introduce reliability where it should not exist. (We note,

FIGURE 16.2

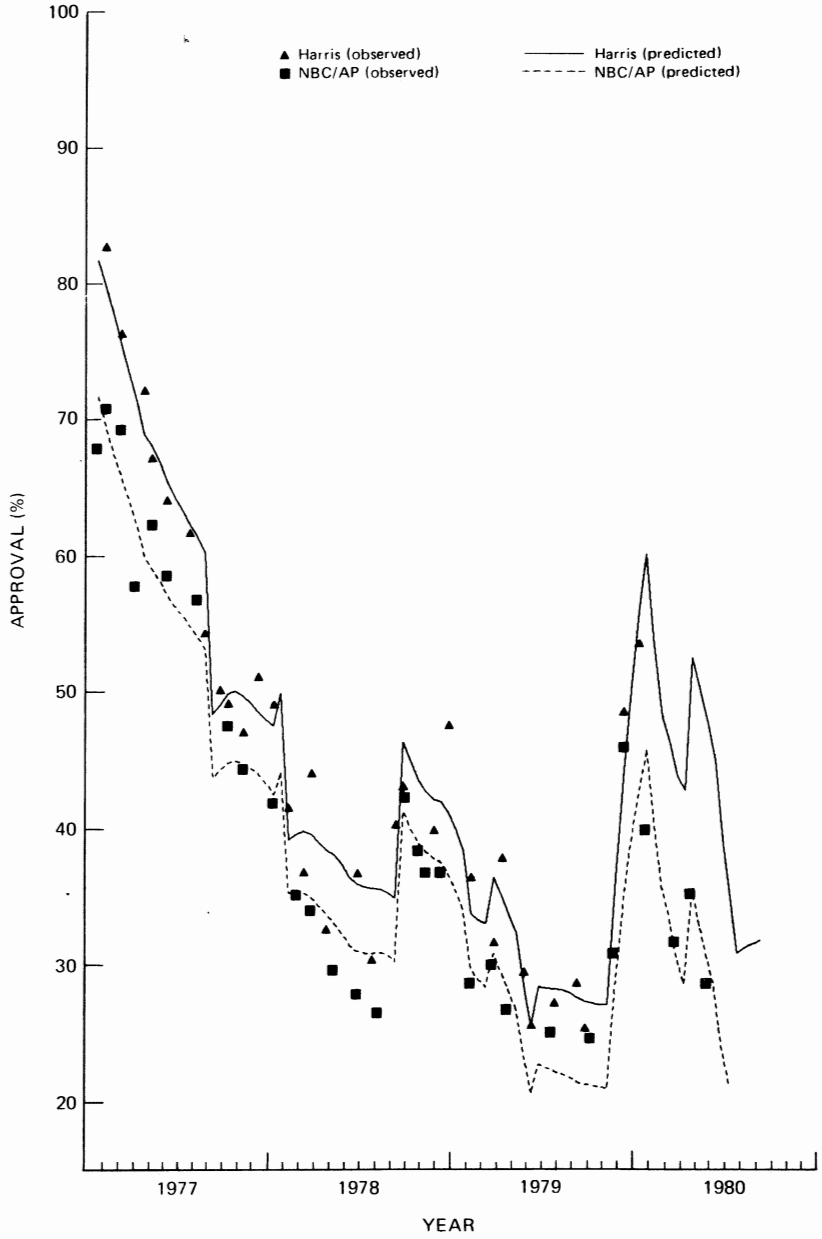
Presidential Approval Measurements Expected Under Custom-Fitted Model (see Table 16.11) and Actual Measurements Made by CBS/NYT, Gallup, and Roper, 1977-80.



NOTE: The time period of this plot is restricted to years in which CBS/NYT made survey measurements.

FIGURE 16.3

Presidential Approval Measurements Expected Under Custom-Fitted Model (see Table 16.11) and Actual Measurements Made by Harris and NBC/AP, 1977-80.



NOTE: The time period of this plot is restricted to years in which NBC/AP made survey measurements.

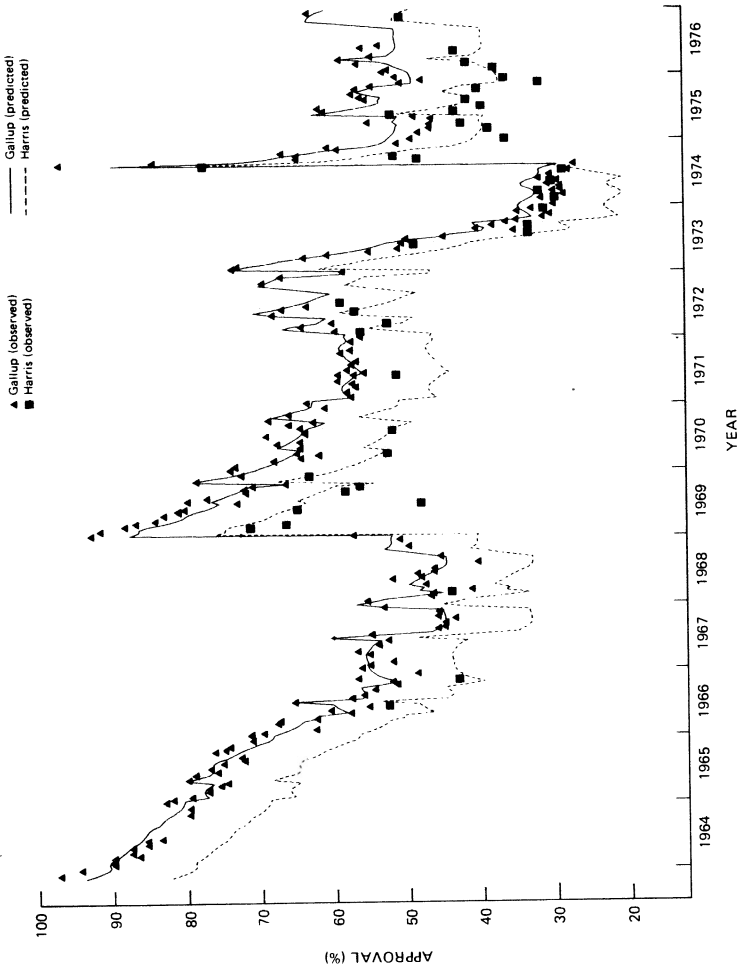


FIGURE 16.4
Presidential Approval Measurements Expected Under Constant-House-Effects Model (see Table 16.8) and Actual Measurements Made by Gallup and Harris, 1963-76.

nonetheless, that one organization [Roper] using a four-category response scale evidenced reliabilities at the same level as those of CBS/*NYT* and Gallup.)

Conclusions

The results of this analysis address two matters: (1) whether the different house/instrument combinations do in fact measure the same thing and thus whether the construct of popularity may be validated in substantive terms, and (2) whether there exist statistically discernible differences between the organizations' precision in measuring popularity.

If survey measurements are to be taken seriously, they must be subjected to external validation (see the panel's recommendations: Volume 1, Chapter 10). Assessing survey organizations' ability to monitor a subjective state, which is in principle unobservable, requires that their measurements be identical, or at least that they diverge in an explicable manner, and that those measurements be related to theoretically motivated observables.

The resolution of the first set of concerns is clear. The variation in all the organizations' measures of presidential popularity may be successfully accounted for with a fairly simple model. It includes economic performance, political events, and a gradual diminishing of each president's inaugural glow. While small distinctions may be discerned in each measure's sensitivity to the political climate, they are all in accord with an expectation based on the nature of the question format. For the most part, however, the readings move in concert. The main dissimilarities between any two organizations' measures seem to be adequately expressed as a constant. The gap between the Gallup-CBS/*NYT* and the Harris-NBC/AP levels appears to be consistent with an explanation centering around the nature of the latter's "only fair" response option. The clearest, and the most heartening, result is that the exact coincidence of the Gallup and CBS/*NYT* questions produces almost identical readings. More troubling is the persistent contrast between the Harris and NBC/AP readings, which are based on very similar (though not identical) question formats. The summary view, though, is one of a set of measures that track a substantively justified presidential popularity in a remarkably coherent fashion.

The concern about instrument precision is a little less easily resolved. It is clear that the Gallup, CBS/*NYT*, and Roper instruments show greater reliability then do the Harris and NBC/AP measures. This divergence was shown both in terms of a direct, substance-free comparison of variability and also in terms of deviations from the predictions of our substantive

models. These separate assessments supplement each other because the first derives from a straightforward statistical test that does not rely on an intricate model specification, while the second takes into account the information produced in the validity test.

The implications of this finding, however, are not self-evident. Should the unreliability of the Harris and NBC/AP readings be due to the organizations' internal practices,[26] a call for firmer operational standards must be heeded if the survey task is to be taken seriously. The same message must be read if the imprecision is due to an undesirable ambiguity in response options; less reliable question formats should give way to more reliable ones. However, if the increased variability reflects a substantive phenomenon, here an attitude uncertainty, which is masked by a simpler format, the tables are turned. Making a judgment about the sources of this difference in instrument reliability requires a more extensive examination of organizational procedures and of individuals' attitudes than can be attempted here. Nonetheless, if Harris and NBC/AP are to avoid being tarred with the brush of unreliability, then further and more extensive research would seem in order.

On a broad scale the picture looks fairly good, at least with regard to assessing presidential popularity using the techniques of survey research. Aside from an obvious constant difference in levels, the instrument distinctions are not large. The differences in the organizations' precision, while requiring action, are of relatively small magnitude compared to the large variations occuring in presidential popularity over time. While measuring popularity requires caution in interpreting the details of the results, our findings do not support claims that such survey measurements are either invalid or entirely unreliable. Rather our analysis suggests two, more specific conclusions.

First, given a large body of data containing frequent measurements of a phenomenon that varies considerably over time, such as presidential popularity, common survey measurement procedures do seem to provide a generally consistent portrayal of the fluctuations that occur in the national population. Survey estimates of these fluctuations, moreover, are systematically related to exogenous variables that are (theoretically) likely causes of shifts in the popularity of presidents. A general claim to reliability and validity for these measurements is thus sustainable.

Second, our analyses also indicate that the range of inconsistencies existing in these survey measurements is sufficiently large that comparison of two measurements made at different times may yield erroneous conclusions about changes in the national population's approval of a president if sampling error is the only source of variance taken into account in making inferences. (This appears to be particularly problematic when the measure-

ments are made by different survey organizations.) Inconsistencies evident in the series we examined suggest that the nonsampling components of variance in these measurements are sometimes as large as are the components attributable to sampling variance. Thus, a cautious analyst may wish to allow correspondingly larger margins of error around point estimates of presidential popularity.

APPENDIX TABLE 16.A
Estimates of Event Impacts

Administration & Event	Date	Estimated Impact
Johnson		
Election 1964	64092 – 64102	0.22
Bomb Vietnam	65021	−3.24
Dominican Republic	65051	4.25
Watts	65082	−0.11
Escalate war	66021	0.70
Demonstrations	66041	1.15
Demonstrations	66051	−4.12
Escalate war	66071	4.91
Urban riots	66072	−7.73*
Election 1966	66101 – 66102	−2.74*
Glassboro Summit	67062	7.60**
Newark-Detroit riots	67071 – 67072	−5.52**
Nonproliferation announcement	67121	7.88*
Christmas message	67122	5.29
Tet Offensive	68021 – 68022	−2.89
New Hampshire primary	68031	3.44
Abdication	68041	2.41
Czechoslovakia invasion	68082	0.08
Election 1968	68092 – 68102	3.10*
Nixon		
Nixon world tour	69022	1.77
Moon walk	69072	1.88
Anti-war moratorium	69102	−4.41
Silent majority speech	69111	13.07**
Cambodian invasion	70051	3.74
State of Union Speech	70012	0.62
Election 1970	70101 – 70102	4.47**
Laotian invasion	71021	−5.04*
Pentagon Papers	71062	−0.90
China visit	72021	5.13**
ITT scandal	72032	−2.17
Mine Haiphong	72051	7.57*
Moscow trip	72052	4.46
Election 1972	72092 – 72102	3.95**
Rebomb Hanoi	72122	−7.79
Peace announcement	73012	15.72**
POW return	73021	2.96
Watergate	73032 – 74081	−1.31**
Brezhnev visit	73062	−2.63
Agnew resignation	73101	−5.60**
John Dean's testimony	73071	−5.38
TV Watergate speech	73082	0.78
Saturday Night Massacre	73102	−2.82
Impeachment hearings	74052	−1.16

APPENDIX TABLE 16.A *(continued)*

Administration & Event	Date	Estimated Impact
Ford		
Nixon pardon	74091	−8.79**
"WIN" speech[a]	74101	1.15
Vladivostok Summit	74112	−3.00
Tax cut	75041	0.81
Mayaguez rescue	75052	12.24**
Assassination attempts	75091 – 75092	2.56*
Fire Nixon cabinet	75111	−4.04*
New Hampshire primary win	76031	0.83
Florida primary win	76032	6.02*
Republican nomination	76081	−26.47
Election 1976	76092 – 76102	6.87
Carter		
MEOW speech[b]	77042	−0.89
Lance hearings	77091	−10.29**
Carter world tour	77122	−0.06
State of Union speech	78012	2.99
Panama Canal Treaty	78021	−9.38**
Camp David Peace Treaty	78092	11.11**
Election 1978	78101 – 78102	−0.16
Deng visit	79021	−3.04
Three Mile Island	79032	3.87**
Gas lines	79052 – 79061	−2.84**
Salt signing	79062	5.44*
Iran Embassy takeover, etc.	79112 – 80012	9.65**
Rescue mission	80042	10.49**
Credit controls	80031	2.14
Billy-gate	80062	−4.24

NOTES: Dates are given in terms of year, month, and whether scored as in the first or second half of month. For example, 79052 indicates the second half of May 1979.

Impact coefficients are in pure percentage points—which is to say that Ford's pardon of Nixon (impact coefficient = −8.79) is estimated to have hurt his standing by about 8.79 points. Most estimates have a standard error of 2 to 4 points and the exact impact is only roughly indicated. In particular, the estimate for Ford's nomination seems overly large and imprecise; its standard error is 75.01. Those events which are individually distinct from zero are so marked. For events taking place over a period of time, the impact was estimated for all included time points and subsequently applied to all. For example, the impact of the Newark-Detroit riots was applied to both the first and second halves of July 1967. Importantly, Nixon seems to have lost about 1.31 points for each half-month of the Watergate period, aside from adjustments for discrete identifiable events.

[a]WIN: Whip Inflation Now

[b]MEOW: (Energy crisis is) moral equivalent of war

*Probability that coefficient is actually zero is less than .10

**Probability that coefficient is actually zero is less than .05.

Notes

1. Two decades earlier, Neustadt (1960:224) noted that presidential-popularity ratings "are very widely read in Washington. Despite disclaimers, they are widely taken to approximate reality."

2. This figure is derived from a review of abstracts of the evening news for the month of June in the years 1973 through 1975 and 1977 through 1979. The abstracts were prepared by the Vanderbilt University Television News Archives (1973–79). A single month was used to make the sampling of stories practical (given our limited resources). For the 6 months in question, we found 13 stories that seemed definitely to include poll data on the president's "popularity" or "job rating," for example, where the abstract read "CBS/*NYT* poll re. Carter popularity." In addition to these stories, we found five more dubious instances, for example, an abstract showing "ABC/Harris poll re. inflation and Carter," and two news reports of poll measurements that might be dubbed "unpopularity" measurements. The latter two stories (CBS and NBC News on June 13, 1974) reported poll findings on public support for the impeachment of President Nixon.

If the five dubious and two "unpopularity" stories were included in our count, we would obtain an average of 3.3 stories per month; if these stories were excluded, the average would be 2.2.

We should also note the conclusions of Paletz et al. (1980:499), who independently coded the CBS and NBC evening news and stories in the *New York Times* for the entire years of 1973, 1975, and 1977. They observed that

> . . . the most frequent [subject for news stories using poll data] we dubbed "presidency"; this category contains assessments and evaluations of the president's job performance and of members of his family. These popularity polls were prominent on television, encompassing almost a quarter of all poll stories, somewhat fewer for the [*New York*] *Times*. At times the pollsters' desire to measure, and the passion of the press to publish, presidential approval ratings appear to be fetishistic. During the first few months of the Carter administration, the *Times* printed three such polls in 24 days . .

3. Marginals and survey dates for the Gallup, Roper, CBS/*NYT*, and NBC/AP measurements were obtained from the Roper Center. Marginals and survey dates for the Harris measurements were obtained from the Louis Harris Data Center (University of North Carolina) for the years 1963 to November 4, 1978. Later dates were supplied by Louis Harris and Associates' New York office.

Surveys were assigned a date corresponding to the midpoint of their field period. For some survey measurements, the archives did not supply precise information on the starting date of the survey or the length of the field period. Often the archival records contain vague notations (for example, early May 1966). We used all available data to estimate midpoints. Where a precise starting date and field period were supplied, the midpoint was computed arithmetically. Where a precise start-date was not known, the following rules were followed: (1) if the "early" part of the month was specified in the archives records, the seventh of the month was coded as the midpoint date; (2) if only the month was specified or if the records specified "middle" of month, the fifteenth of the month was coded as the midpoint; (3) if the "late" part of the month was specified, the twenty-first was coded as the midpoint; (4) if the survey was said to span 2 months, but no precise dates were given, the last day of the first month was called the midpoint; (5) if the survey occurred in January of a year in which a new president was inaugurated, a survey measuring approval of the new president was assumed to have a midpoint of January 25.

4. We do note, nonetheless, that if Organization 1 consistently had longer field periods (than Organizations 2 and 3) during times when public opinion was shifting rapidly, the agreement between Organizations 1 and 2 and 1 and 3 would be expected to be less than that between Organizations 2 and 3.

The length of field periods for the surveys in our database averaged 4.4 days for Gallup, 5.4 days for Harris, 3.8 days for CBS/*NYT*, 2.1 days for NBC/AP, and 8.7 days for Roper. In our analysis we found (see Table 16.5) that the highest correlation (+0.986) occurred between the organization with the longest field period and that with the second shortest—suggesting that similarity in field periods is probably not a dominant factor in explaining the patterns of agreement between these measurement series.

5. The time periods were constructed by beginning from January 1 of each year and dividing a decimal representation of the year into 20 equally spaced units. Time periods were subdivided into separate periods if there was a change of administrations.

6. Specifying the factors that produce this noise and estimating their effects remains a challenging task for future analysts. To facilitate their work we have deposited a copy of our dataset with the Roper Public Opinion Research Center (P.O. Box 1732, Yale Station, New Haven, CT 06520).

7. A weighted analysis of these approval ratings produced coefficients that varied by no more then 0.01 from the (unweighted) coefficients shown in Table 16.5. The weighted analysis was designed to place more weight upon comparisons involving large samples. Weights were calculated for every pair of organizations in each time period. (Thus there were 61 separate weights for the Gallup-Harris comparison—one for each time period in which both organizations made measurements.) Each weight was the reciprocal of a (upper bound on the) standard error for the expected difference between estimates of proportions based upon samples of the size (N_1, N_2) used by the organizations in that particular time period $[(.25/N_1) + (.25/N_2)]^{0.5}$. (It should be noted that samples in question may sometimes be the sum of the samples in two surveys conducted by an organization in the same time period.)

8. The incompleteness of our 3-way table introduces difficulties in the computation of the degrees of freedom for the chi-square statistics. As Haberman (1979:469) notes in his text: "Unfortunately, no simple rules for degrees of freedom appear to be entirely adequate when the generating class of the hierarchical model has at least three members."

Because of this difficulty, we note explicitly the algorithm we used to calculate the degrees of freedom in Table 16.6. Let $M_{ij} = 1$ if organization i made a popularity measurement in time period j; otherwise $M_{ij} = 0$. (Remember that our 100 time periods include only instances in which two or more organizations made measurements.) We calculated the degrees of freedom for each model as:

$$\text{Model 1: } d.f._1 = \left(\sum_{j=1}^{100}\sum_{i=1}^{5} M_{ij}\right) - 1$$

$$\text{Model 2: } d.f._2 = \sum_{j=1}^{100}\left(\sum_{i=1}^{5} M_{ij} - 1\right)$$

$$\text{Model 3: } d.f._3 = \sum_{i=1}^{5}\left(\sum_{j=1}^{100} M_{ij} - 1\right)$$

$$\text{Model 4: } d.f._4 = d.f._3 - 5 + 1$$

Allowing for some uncertainty in the calculation of d.f., we observe that the L^2 values in

Table 16.6 are so large that an error in the reported d.f. would not have a meaningful impact on our inferences about the fit of Models 1 through 4.

9. The mean absolute discrepancy was computed as:

$$\left(\sum_{i=1}^{N} |O_i - E_i|\right)/2N,$$

where O_i is the observed frequency count at time i, and E_i is the count expected under the model. N is the total number of residuals (which is equal to the number of nonempty cells in the table of observations). This measure has sometimes been termed the index of dissimilarity. It indicates the percentage of observations that would have to be changed to bring the observed data into perfect agreement with the values expected under the model.

10. To perform these analyses, we altered the normal execution of the ECTA program (Goodman and Fay, 1973) in order to write the table of observed and expected values for our models onto a reusable storage medium. The resultant dataset was then reanalyzed using a special-purpose program designed to provide the measures of residual dispersion presented in Table 16.7. (The residuals from the fitting of the various substantive models (see below) were subject to a similar reanalysis to provide the measures of residual dispersion shown in Tables 16.12 and 16.13.)

It should be recognized that the decomposition presented in Table 16.7 was obtained by computing the *actual* contribution of the measurements made by each organization to the likelihood-ratio chi-square for Model 4 of Table 16.6. This is an exact arithmetic decomposition, i.e., the sum of each organization's contributions equals the total chi-square for Model 4. A somewhat different procedure might also be applied to these data; this procedure requires us to compute models constrained to fit all data points except those produced by organization 1 (then 2, etc.). The difference between the fit obtained in the overall analysis and that obtained when the measurements produced by organization 1 are "blanked out" can be taken as another indicator of the contribution of that organization's measurements to the poorness of fit of the overall model. This method does not, however, provide an exact decomposition (i.e., the sum of each organizations "contributions" does not equal the chi-square for the overall test). This occurs because slightly different models are being used in each case (due to the fact that a slightly different set of data points is being fit in each instance). For the present data, such an analysis yields likelihood-ratio chi-squares of: 1,452.1 (overall), 302.7 (without Gallup), 443.7 (without Harris), 1,407.8 (without CBS/*NYT*), 1,209.4 (without NBC/AP), and 1,283.8 (without Roper). Subtracting, we would obtain estimated contributions under this procedure of: 1,149.4 (Gallup), 1,008.4 (Harris), 44.3 (CBS/*NYT*), 242.7 (NBC/AP), and 168.3 (Roper).

11. Setting the initial conditions normally would have a noticeable effect on subsequent values. However, by setting the starting point as though the process were in equilibrium in January 1961, and by running the process for two years before any estimate was made, the effects of the starting conditions virtually wash out. The only relevant impact of the initializing input was that due to inflation, and its weight by January of 1963 was only 0.000409. All other components have a natural starting point at a subsequent time.

12. The time-stream was divided into half-months (beginning on the first and the sixteenth of each month). Thus the time scale is roughly comparable to that of the first set of analyses (one-twenty-fourth rather than one-twentieth of a year); a very minor imprecision is introduced for February. As before, coincident readings by the same house were given a weighted average for that half-month.

13. This substantive model is based on earlier work by Mueller (1970, 1973), Kernell (1978), Kernell and Hibbs (1981), Stimson (1976), and Monroe (1978). The examination of movement in presidential popularity is now a bit of a cottage industry. The war and economic variables are nothing new at all. Allowing the "glow" or honeymoon decay to be estimated is slightly novel, and the results suggest that this effect persists somewhat longer than was previously thought. The event series are a substantive addition. They were found to be necessary to obtain stable estimates for the other elements, a finding that suggests their place in a full specification. Response to the economic variables here are modeled as a constant process across all administrations. Some experimentation suggests that this might be an oversimplification, as might be expected if presidents are differentially judged by the emphasis they give to different national problems.

14. The actual level of unemployment was used here rather than a change rate because (1) that is the common reference point in popular discussion, and (2) experiments with other formulations suggested no stark differences. Furthermore, the measure of unemployment focused on adult males so as to control for seasonal and long-term changes in labor force composition. (Finding such a low and imprecise parameter for unemployment suggests that more substantive work may be useful here. See previous note.) The inflation rate used is the percentage change from the previous time point's Consumer Price Index and thus is not seasonally adjusted. Because the time-series is defined in terms of half-months, both economic series represent linear interpolations for the rate on the seventh and twenty-first days of each month. The actual number for inflation is multiplied by 24 so as to produce an approximation of the familiar annualized metric for each half-month.

15. At first a sequence of 69 variables was generated (see Appendix 16.A for details), all zero except for a unit spike to mark the fortnight during which a significant event occurred that was thought to bear on presidential popularity. (The inclusion of particular events is a matter of judgment, and the strategy adopted here was to be liberal in allowance. *The New York Times Index* and *The World Almanac* provided the dates of the relevant events.) Then, given a single dynamic parameter (which was determined by iterating on this process), the magnitude and a sign for each event was estimated with OLS (ordinary least-squares). Inspection of the time paths suggested that some especially significant events had a longer-lasting impact—the Newark-Detroit riots, the Saturday Night Massacre, John Dean's testimony on Watergate, Camp David, and the several presidential election campaigns—and so another event series with a larger d (a longer lag structure) was designed for them.

It should be noted that some error variance might be mistakenly attributed to the occurrence of specific events, but worries about merely matching up "white noise" ought to be alleviated in that (1) the estimate for each event is constrained because a single reequilibration parameter was used for all events, and (2) of the events whose individual impacts were statistically discernible from zero, every one has the correct sign (something definitely not expected if those terms mostly represented sampling error). In any case, our purpose was to eliminate from the error term any movement that might plausibly be attributed to real change rather than to measurement error.

The event magnitudes estimated by this procedure are presented in Appendix Table 16.A. After the rough magnitude of the event impacts was estimated, a single pair of series of political events was created, for the reevaluation of the dynamic parameters and the house/question differences, by simply replacing each of the unit spikes with its estimated magnitude. The dynamic parameters were then estimated using a nonlinear procedure. Then the entire process was repeated until reasonable convergence was obtained. (This back-and-forth process could have been accomplished in one large, nolinear-equation estimation, but the cost was practically prohibitive.)

16. A model, transforming the popularity proportions into logits, produces a similar (though slightly less satisfactory) fit and pattern of results. Using a linear scale for popularity makes it difficult to predict the initial, high values of popularity in each administration; the logit transformation exacerbates that problem. More elaborate transformations might improve the precision of the model and reduce the estimated error components, but given the relatively small number of bad misses, should not change the substantive conclusions much. The nominal scale of percentages is here maintained to ease interpretation.

17. This procedure depends on the assumption that the additional Gallup "approvers" come only from the Harris "only fair" category and that all Harris approvers are also Gallup approvers. The *Post* data suggest this is only slightly unrealistic. The percentages giving a Gallup "approve" coming from each of the Harris categories are: "excellent," 7.45; "pretty good," 61.37; "only fair," 29.18; "poor," 0.6; "don't know" 1.4. The size of the "only fair" contribution suggests the possible importance of the difference between Gallup and Harris "approvers." Only 3.7 percent of Harris "excellent" and "pretty good" group did not also say "approve" to the Gallup question.

18. The Gallup-Harris comparison used here takes advantage of the greater range of change covered by those two houses' time spans. The analysis is limited to the last three administrations because Harris provided too few Johnson administration readings ($N = 3$) for precise measurement. Other Harris samples during the Kennedy and Johnson years were of special (voter) samples, and they were dropped from the analysis because such samples might manifest different responsiveness that was due to the intentionally special character of their respondents rather than to the survey procedure.

19. Allowing the variables' effects to be different for each house, as was done for the equation in Table 16.10, indicates that the NBC/AP and Harris measures are not significantly different in their substantive responses. This is in accord with the results of Table 16.11. However, the constant effect (of NBC/AP's rating being lower) persists.

20. Finding that the NBC/AP readings were lower is not altogether surprising. Their question refers to the president by name only, not as "President Carter." In addition, the response category of a "good job" may be more difficult for uncertain supporters to choose compared with Harris's "pretty good." The magnitude of this difference, however, is troubling; the wording dissimilarities are fairly minor. Moreover, whether the constant divergence in measurement level is due to the differences in question wording, or to the organizations' internal sampling and interview procedures, cannot be determined from these data.

21. The mean absolute value of the residuals was chosen rather than, say, the standard deviation, because the inability of a house to match the prediction may be as much due to the model's inadequacy as to the house's shortcomings. This consideration suggests that large deviations should be weighted proportionally to their absolute value rather than to their squared value, so as not to unduly penalize a set of predictions as a consequence of a single, large miss.

22. The slight reduction of the Gallup accuracy reflects the diminished proportion of Gallup readings represented in the number of readings to be estimated. Thus, the estimation equation has fewer Gallup errors to be minimized.

23. The parameters for each substantive variable are mostly statistically undifferentiable across houses, though they do wobble about a little so as to maximize the fit for each particular series. On the one hand, finding differences is not troublesome because the purpose is to

measure deviations off some (not precisely specified) function of the external environment, and no single house has a claim on validity (as the Gallup series would have to in order to rely on Table 16.12's Gallup-dominated numbers). The problem lies in each separate model's exhausting 6 degrees of freedom, a substantial portion of the CBS (20) and Roper (18) observations. Making the theoretical statistical correction, however, yields the same rough rank-order, though CBS falls closer to the Harris-NBC end of the spectrum. The mean squared error (a variance rather than absolute deviation measure) for the houses read 6.64, 12.73, 10.42, 13.18, and 4.07 for Gallup, Harris, CBS/*NYT*, NBC/AP, and Roper, respectively.

24. An expectation may be derived by taking the standard sampling error for *random* samples ($[pq/(n-k)]^{.5}$) of a dichotomy and multiplying it by the expected value of (twice) the right-hand side of the normal distribution. This, of course, assumes that the errors cumulate a number of small disturbances and thus are distributed approximately normally. Mechanically, this estimation took the sample for each survey and the observed approval as the true proportion, calculated the expected deviation due to sampling, and then averaged across all observations. The expectations estimated by this procedure for the Carter years are as follows: Gallup, 0.91; Harris, 0.96, CBS/*NYT*, 1.05; NBC/AP, 0.92; and Roper, 0.88. The differences are due to some distinctions in the times when ratings were taken (the estimated proportion differed) and variability in the sample sizes of each house.

To be sure, these organizations do not take simple random samples, and the true error must be somewhat larger, but this number provides a theoretically motivated benchmark. Differences between the average absolute deviation expected by random sampling and the observed absolute deviations from the estimated model are these: Gallup, 1.28; Harris, 1.72; CBS/*NYT*, 0.75; NBC/AP, 1.63; and Roper, 0.41. Understanding that the actual sampling error is somewhat larger suggests that relative differences in the magnitudes of these "measurement errors" may be even larger

25. The slight reliability advantage NBC/AP has over Harris may be due to the wider semantic gap between "good" and "only fair" as opposed to "pretty good" and "only fair."

26. O. D. Duncan suggested to us another source for cross-house differences—one that is distinct from the quality of instrumentation. It may be that the timing of the surveys is itself a function of matters that may generate measurement variance. For example, some houses may deliberately attempt to catch turning points in popularity (which are presumably newsworthy), and thus find they are trying to measure public opinion when it is unusually and genuinely unstable. For this possibility to account for the observed pattern of measurement variance, it would have to be the case that NBC/AP and Harris were more inclined (than CBS, Gallup, and Roper) to send their interviewers out in stormy times.

References

Brody, R. A. and Page, B. I. (1975) The impact of events on presidential popularity: the Johnson and Nixon administrations. In A. Wildavsky, ed., *Perspectives on the Presidency*. Boston: Little, Brown.

Converse, P. E. (1979) The Impact of Polls on National Leadership. Paper presented at Symposium on the Fiftieth Anniversary of the Social Sciences Research Building at the University of Chicago, Dec. 16-17.

Drew, E. (1979) Reporter at large: phase in search of a definition. *New Yorker* (Aug. 27):45-73.

Goodman, L. A. (1968) The analysis of cross-classified data: independence, quasi-independence, and interactions in contingency tables with or without missing entries. *Journal of the American Statistical Association* 63:1091-1131.

Goodman, L. A. (1970) The multivariate analysis of qualitative data: interations among multiple classifications. *journal of the American Statistical Association* 65:226-256.

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

Goodman, L. A. (1978) *Analyzing Qualitative/Categorical Data*. Cambridge, Mass.: Abt Books.

Goodman, L. A., and Fay, R. (1973) ECTA program: description for users. Unpublished ms., Department of Statistics, University of Chicago.

Haberman, S. J. (1978/1979) *Analysis of Qualitative Data*. Two volumes. New York: Academic Press. (Volume 1, 1978; Volume 2, 1979).

Kernell, S. (1978) Explaining presidential popularity. *American Political Science Review* 72:506-522.

Kernell, S., and Hibbs, D. (1981) A critical threshold model of presidential popularity. In D. Hibbs and H. Gassbender, eds., *Contemporary Political Economy*. Amsterdam: North Holland.

Koyck, L. M. (1954) *Distributed Lags and Investment Analysis*. Amsterdam: North Holland.

Martin, E., McDuffee, D., and Presser, S. (1981) *Sourcebook of Harris National Surveys: Repeated Questions, 1963-1976*. Chapel Hill, N.C.: Institute for Research in the Social Sciences (University of North Carolina).

Monroe, K. (1978) Economic influences on presidential popularity. *Public Opinion Quarterly* 42:360-369.

Mueller, J. (1970) Presidential popularity from Truman to Johnson. *American Political Science Review* 64:18-34.

Mueller, J. (1973) *War, Presidents and Public Opinion*. New York: Wiley.

Neustadt, R. E. (1960) *Presidential Power: The Politics of Leadership* New York: Wiley.

Orren, G. (1978) Presidential popularity ratings: another view. *Public Opinion* May:35.

Paletz, D. L., Short, J. Y., Baker, H., Campbell, B. C., Cooper, R. J., and Oeslander, R. M. (1980) Polls in the media: content, credibility and consequences. *Public Opinion Quarterly* 44:495-513.

Sigelman, L. (1981) Question order effects on presidential popularity. *Public Opinion Quarterly* 55:199-207.

Stimson, J. A. (1976) Public support for American presidents. *Public Opinion Quarterly* 40:1-21.

Sussman, B. (1978) Jury is still out on Carter: distortion in popularity polls. *Washington Post* Feb. 12:A1.

Theil, H. (1971) *Principles of Economics*. New York: Wiley.

Vanderbilt University Television News Archives. (1973-79) *Television News Index and Archive*. Nashville, Tenn.: Vanderbilt University Television News Archives-Joint Universities Library.

Von Hoffman, N. (1979/1980) Public opinion polls: newspapers making their own news? Syndicated by King Features, Inc., July 30, 1979 (Reprinted in *Public Opinion Quarterly* 44:572-573).

Acknowledgments

We are grateful to Theresa DeMaio who collaborated with us at the outset of this work. We also wish to thank Clifford Clogg, Otis Dudley Duncan, Robert Fay, Stanely Presser, and Tom Smith for their helpful advice. (The order of authorship is alphabetical.)