What's Your Temperature?

Thermometer Ratings and Political Analysis

Nicholas Winter
Center for Political Studies
University of Michigan
nwinter@umich.edu

and

Adam Berinsky
Department of Politics
Princeton University
berinsky@princeton.edu

August 1999

Paper prepared for presentation at the Annual Meeting of the American Political Science Association, Atlanta, GA. We would like to thank Chris Achen for his helpful advice. Both authors contributed equally to the paper and order of the names was determined at random.

I. Introduction

In 1964, the National Elections Study (NES) introduced the "feeling thermometer" measures as a way to gauge respondents' affect towards prominent political groups and figures. Since then, thermometer scores have become regular players in political behavior research. These items have been used in scores of studies as measures of a wide variety of concepts, including group affect, racism, sexism, candidate vote intentions and comparative evaluations, candidate affect, and various aspects of gender and racial identification and consciousness, among others.

In most cases, the thermometer score rating of the relevant political person or group has been treated rather unproblematically as an error-free, continuous measure of the concept of interest. On its face, such an approach appears reasonable. The task entailed in answering the thermometer score, after all, seems simple; respondents are asked to rate political objects on a 101-point scale.[1] But though this undertaking appears straightforward, little research has been carried out to see exactly how respondents actually use the feeling thermometers. And what work has been done is not encouraging. Brady (1985) suggests that feeling thermometer scores may be plagued by problems of inter-personal incomparability – that is, different people may interpret the scale in different ways, thereby prohibiting the comparison of scores across individuals. This view has been supported by the work of Green (1988) in the realm of liberal/conservative and Democratic/Republican feeling thermometer, and more generally by the work of Wilcox, et al. (1989). But these works stand as exceptions. Given their widespread use in the political science literature, we know surprisingly little about the measurement properties of feeling thermometers. In this paper we begin to fill this gap in the literature.

We start with a consideration of a basic, often unspoken, assumption about thermometer score; that such scores are interval level data. We then examine how systematic differences in the way that individuals

[1] Additionally, respondents are told that "ratings between 50 degrees and 100 degrees mean that you feel favorable and warm toward the person (group). Ratings between 0 degrees and 50 degrees mean that you don't feel favorable toward the person (group) and that you don't care too much for that person (group). You would rate the person at the 50 degree mark if you don't feel particularly warm or cold toward the person (group)."

use the feeling thermometer scores – what Brady terms inter-personal incomparability – may create problems in our analysis. We conclude with some preliminary recommendations concerning the general use of feeling thermometers. Our analysis will focus on group thermometer ratings, although we will make some reference to ratings of individual figures as well.

II. Faux Interval Scale?

The most obvious concern about the feeling thermometer relates to people's inclination and ability to make meaningful use of a scale with 101 distinct points. Ultimately, people are probably not capable psychologically of making such fine distinctions of their own internal dispositions, even if they could be considered to have such fine-grained opinions to begin with. Moreover, if respondents answer survey questions from the "tops of their heads," (Zaller and Feldman 1992; Zaller 1992) they are unlikely to translate their internal affect toward a group or an individual onto the scale with anything like the precision implied by the 101-point scale. It is almost certain that feeling thermometer scores convey nothing near the exactness implied by the scale.[2]

This is not to imply that feeling thermometers are worthless, but we are left with the question of exactly how much information about the people's internal psychic states is conveyed by a thermometer score ranking. Clearly thermometer scores contain some information, since they pass the basic face validity criterion of correlating appropriately with each other and with other political variables (for example: in the 1992 NES, T-scores[3] of the Democratic party and of liberals correlate 0.46; and the ratings of the Democratic party and the Republican party correlate with party identification 0.64 and –0.55, respectively). There is, however, a continuum of nuance that this information may take. At one extreme, thermometer scores could really be an interval-level measure, analogous to the actual thermometer they are modeled on. Insofar as thermometer rankings approach this interval-level ideal, they convey information about rank order of preferences among groups, as well as information about the degree of difference in preferences. At the other extreme, the

[2] This is somewhat analogous to recording length to the nearest micron using an ordinary tape measure. So, we concede that the measure is not *that* precise, but on the other hand it is not clear that it needs to be.

[3] We use the term "T-score" as a synonym for "thermometer score," for aesthetic relief.

thermometer scores may convey only the simplest ranking information. If this is the case, then we can say with confidence that a respondent prefers one group to another if she rates it higher, but we can say no more. Imagine, for example, that I rate environmentalists at 60, liberals at 70, and the Democratic party at 90. Under the latter (rank only) model, this indicates that I like Democrats more than I like liberals and that I prefer both to environmentalists. Under the interval-level model, these scores not only indicate that I feel relatively warm toward all three groups (since I rate them all above 50), but also that the difference in my affect toward environmentalists and liberals is about half the difference in my affect toward Democrats. Of course, the truth lies somewhere between these extremes. But where?

To answer this question, we first examined how many unique values respondents used in their 27 ratings of groups and their 16 ratings of individuals. The number of unique points that a respondent uses on the 101-point scale sets a lower limit on the amount of information their responses can convey: if someone rates some individuals at 0 and others at 100 (and uses no other rankings), for example, all we know is that she prefers the latter group to the former group. If, on the other hand, she rates the individuals with 16 distinct rankings, there is the possibility that information beyond mere rank is being conveyed.[4]

In order to avoid ceiling effects, this analysis includes only respondents who rated at least half of objects presented (14 of 27 groups; 8 of 16 individuals).[5] When rating groups, the average respondent used about seven unique points on the thermometer scale (mean=6.8; sd=1.9), and about 90 percent of respondents used between four and nine points, inclusive. For ratings of people, the average respondent used about six unique points (mean=6.2; sd=1.7), and over about 93 percent used between four and nine points.

[4] Of course, the number of unique values does not tell us directly how much information is being conveyed. On the one hand, someone who ranks many groups at the same point on the scale could be telling us – with great precision – that she really does feel the same warmth toward them all. On the other hand, the use of many distinct rankings need not correspond to more information if there is random measurement error. However, the number of unique values does set a floor on the information that is actually *conveyed*, and can give us an initial view of how people use the scale.

[5] For the group thermometers, about 90% of respondents (2,225 of 2,485) rated passed this test. Of the remaining 243 respondents, the vast majority rated only the first two groups in the battery – the Democratic and Republican parties – and then declined to rate all subsequent groups. For ratings of individuals, 96% of respondents (2,376 of 2,485) passed the test.

This suggests that most respondents, at least, are making reasonably nuanced evaluations that distinguish among at least among four levels of warmth.

If we combine ratings of group and individual objects, however, the average respondent (of those who rated at least 22 of 43 objects) used 13 unique scale points (mean=13.0; sd=3.0). This seems to be a specific case of a general pattern. Respondent tend to use more unique scale values the more objects they rank – each additional four items they rank increases the number of unique values they use by about one.[6] So, this suggests that that individual respondents do *not* have a small set of T-score values – such as 0, 50, and 100; or 25, 50, 75 and 100; or whatever – that they use to rate all objects. Instead, as respondents are called upon to rate additional objects, they seem to use additional points on the scale in order to make finer distinctions among the various objects.

What we still do not know is how much information is conveyed by the magnitude of the differences among the thermometer scores. In other words, how much real information is there in the actual temperatures that people choose? Consider, for example, the difference between my rating of liberals and my rating of conservatives. Let us assume that I gave liberals a 40, and conservatives a 60. We know, at least, that I prefer conservatives to liberals, and we would expect, therefore, that I would identify my own ideology as conservative as well. Imagine that you rate liberals at 10, and conservatives at 90. You, too, have a preference for conservatives, and we would expect you also to self-identify as a conservative. If thermometers convey *only* rank-order preference information, then that is all we can say. To the extent that the thermometers scores contain interval-level information, however, we would expect the difference between our liberal and conservative thermometer scores to relate more systematically to ideological self-placement. You would be more likely then am I to self-identify as a conservative, and you should place yourself further in the conservative end of the ideological scale.

This idea underlies the next set of analyses, in which we plot differences in oppositely-valenced thermometers (Democrats-Republicans; Bill Clinton-George Bush; liberals-conservatives; and blacks-whites)

[6] In the bivariate regression of number of unique values on number of objects ranked, b=0.27, se=0.01, p<0.001.

against other NES measures that should, theoretically, be related to the respondent's difference in affect toward the two objects. Figure One shows, in stylized form, the patterns we might expect to find under the two extreme cases discussed above – the interval case and the ordinal case. The straight line corresponds to the case where thermometer scores contain ideal interval information: the mean value of the indicator variable increases linearly as the difference between the two thermometer scores increases. The step function is a stylized version of the case where thermometer scores convey only rank order information. In this case, on average, those individuals who rate one group significantly higher than another are very different on the indicator variable from those who rank the two groups in an opposite manner. But thermometer differences *within* each of these two groups of individuals do not map perfectly to changes in the level of the indicator variable. To illustrate the comparison between the ordinal and interval stylized cases with an example, imagine that I rate Democrats at 60 and Republicans at 40 and you rate Democrats at 70 and Republicans at 30. If feeling thermometers are truly interval-level data, we would expect you to score higher on the indicator variable – say affect toward the two parties – than I do, because your thermometer difference is higher. Under the ordinal case, however, we would expect equivalent scores on the indicator variable; all the thermometer score tells us in this case is that we both prefer Democrats to Republicans.

We look at seven such comparisons. First, we compare the party thermometer difference with a summary of the number of things respondents mentioned that they like and dislike about the parties.[7] We expect the likes/dislikes measure to reflect the polarization of respondents' feeling about the two parties fairly well; the question is the degree to which the thermometer score difference predicts this polarization. The answer, in Figure Two is that the thermometer predicts this difference reasonably well.[8] The line is relatively straight, with some flattening on the ends. This indicates that the thermometer difference does indeed carry

[7] This measure was based on the NES battery that asks respondents for up to five things they like and dislike about each party. The measure was taken by adding together the number of Democratic likes (v923402-v92346) and Republican dislikes (v923420-v923424), and subtracting the number of Republican likes (v923414-v923418) and Democratic dislikes (v923408-v923412).

[8] For these graphs, respondents were grouped into 21 groups: those for whom the thermometer difference was between 95 and 100; 85-95; and so on. The average value of the indicator variable (the likes/dislikes summary in this case) for each of these groups was then plotted against the thermometer difference.

significant information not merely about respondents' rank ordering of the parties, but also about their *degree* of affect toward the two parties.[9] The next comparison, between the party thermometer difference and mean partisanship, tells a somewhat different story. Here, the line is distinctly curved (Figure Three). The thermometer difference predicts partisanship, but most of the action is in the area around a zero difference. Once a respondent rates one party higher than the other by 30 or 40 points, she is extremely likely to identify with that party; additional difference in the ratings does not correspond to significantly stronger partisanship.

Turning next to candidate evaluations, in Figure Four, we compare the difference in candidate thermometers (Clinton minus Bush, using pre-election thermometers) with a summary measure of likes and dislikes about the two candidates.[10] This figure is quite straight – again indicating that the thermometer difference is essentially linearly related to how respondents feel about the candidates, based on how many positive and negative things they have to say about the two candidates. Figure Five, which makes the comparison between the post-election thermometer difference and Clinton's portion of the two-party vote, is strikingly different; the candidate feeling thermometer difference has a curved relationship with vote choice. This relationship is perfectly reasonable and reflects the lack of information in the vote-choice measure.[11] Essentially, people who prefer one candidate over the other even slightly (as measured by thermometer scores) are overwhelming likely to vote for that candidate. Because the vote measure does not include any information about the *degree* of preference for the candidate, and because vote corresponds closely to preference (as we would hope it would!), the line rapidly approaches the theoretical limit of zero or one.[12]

Next, we compare the difference in the ratings of liberals and conservatives with ideological self-placement (Figure Six). These figures again suggests a highly linear relationship, with differences in

---

[9] Beyond a difference of roughly 60 points there is a ceiling effect as the lines flatten out. This could indicate two things: first, that beyond a 60 point difference, the thermometer score does not really correspond to more positive affect; or second, that although larger differences in thermometer do correspond with stronger affect, that respondents are unable to think of additional things to mention that they like or dislike about the parties – a ceiling effect of sorts.

[10] This measure was constructed analogously to the party likes and dislikes measure, using variables v923110-v923114 (Bush likes), v923116-v923120 (Bush dislikes), v923122-v923126 (Clinton likes) and v923128-v923132 (Clinton dislikes).

[11] The vote choice variable is v925608.

[12] Two measures of the *degree* of support for the major parties – non-voting and voting for Perot – are both predicted by *low* thermometer score differences in a reasonably linear way. This is also consistent with thermometer scores containing some degree of interval information.

thermometer ratings of the two ideological groups corresponding quite closely to average ideological identification. The only oddity is the dip at plus 50.

Finally, in Figure Seven and Eight we compare the black-white thermometer difference with two different measures of racism: racial resentment,[13] and endorsement of stereotypes about blacks.[14] This figure was prepared using only white respondents. Both figures tell the same story. On the one hand, the lines are relatively straight, indicating that the thermometer difference does contain information about the *degree* of whites' dislike toward blacks. On the other hand, however, the lines are quite flat, and there does not appear to be a strong relationship between the thermometer score difference and the two racism measures. Looking at the frequencies makes clear why this is the case: almost half (49 percent) of whites give blacks and whites identical temperature scores, so this measure has significantly less variance than the other thermometer differences we have examined.[15] But the bottom line remains the same. The analyses here, like those presented above, indicates that the feeling thermometer contains at least some interval-level information.

Summary

Based on our assessment of the interval nature of thermometer scores, we can come to some initial conclusions. First, while respondents clearly do not make real use of the 101 points available on the thermometer scale, they do make use of a reasonable number – between four and nine in the vast majority of cases. This suggests that the thermometer is conveying as much information as the five- and seven-point scales that appear on myriad other NES questions. Moreover, there is circumstantial evidence that as respondents encounter additional thermometers, they make use of additional scale points in order to make finer distinctions among the groups they are rating. This result is promising because it suggests that

[13] Racial resentment, a measure of symbolic racism, is calculated by averaging four variables: v926126-v926129 and recoding to the zero-one interval. See Kinder and Sanders (1996, pp. 106-115) for details on the measure.

[14] This measure is the difference between respondent's assessment of blacks as lazy, unintelligent, and violent; and their assessment of whites on those dimensions. Variable numbers v926221-v926222, v926225-v926226,and v926229-v926230.

[15] Twenty percent of respondents rate Democrats and Republicans equally; 26 percent rate liberals and conservatives equally; and 12 percent (pre) and 11 percent (post) rate Clinton and Bush equally. While the large number of whites who rate blacks and whites equally may be a reflection of the happy decline of negative feelings toward blacks in American society, we believe that at least some of the respondents that rate them equally are doing so to hide negative feelings toward blacks. This is a possibility that we plan to investigate more fully in later work.

respondents are paying attention to the task at hand when they answer the thermometer questions – they are not are simply rattling off a series of unconnected numbers to satisfy the demands of the survey interviewer. Finally, the evidence suggests (with the exception, perhaps, of the party ratings[16]) that the thermometer scales are conveying reasonably nuanced "interval-like" (if not true interval-level) information. In sum, then, it seems reasonable to treat the feeling thermometers as interval-level scales.

III. Inter-Personal (In)Comparability?

But establishing that people make meaningful use of the feeling thermometer scales is only the first step. Even if respondents make meaningful use of the thermometers, different people may use the thermometer scales in different ways. Theoretically, we may expect this to be the case. The findings reported in the last section suggest two possible models for how respondents go about answering the thermometer questions. Unfortunately, both these models raise serious questions of inter-personal comparability. First, when prompted for how they feel about a group, respondents may well begin by deciding if they feel positively, negatively, or neutral toward a group. They may then modify that judgement with a "strongly" or "mildly" (i.e. "strongly negative," "mildly positive," "just positive of neutral," etc.). These two steps would generate an implicit three-, five- or seven-point judgement assessment. Finally, respondents would translate their assessment onto the 101 point scale; deciding, for example, that "strongly negative" corresponds to 10. This stage would almost certainly create problems with interpersonal comparison of thermometer scores, since different people would convert their judgements into temperatures differently.

A second model possible model for the thermometer response involves a sort of "anchoring and adjustment" process. Under this model, respondents calculate their response to each item after the first in terms of their *relative* feelings toward that object, compared with what has come before. This model suggests that each respondents' thermometer pattern is path dependent, in the sense that their scores will depend in

[16] One possibility that would explain why the party thermometers display a less interval-like pattern has to do with people's psychological identification with one or the other party. If people identify with the Democrats or Republicans, this may lead them to give a simple "I like them" or "I hate them" response when asked to rate one of the parties; whereas for other groups they may think more carefully about their feelings.

part on how they rated earlier groups. The set of scores for a given individual respondent would convey rather precisely how they feel about the different groups in relation to each other, but their particular scores may bear no comparison with those of other respondents. In both cases, the thermometer scores may not be compared across individuals in a meaningful way.

Whatever model people are using, then, we have reason to believe that different people may be using the thermometer scale in different ways. If so, this can lead to serious inferential difficulties in statistical analysis that makes use of T-scores. As Brady (1985) has shown, interpersonal incomparability in the use of any interval-level scale can seriously distort measures of association – such as correlation and regression coefficients – in bivariate analyses. Work by Green (1988) suggests that such incomparability pervades feeling thermometer scores. As he writes, "the feeling thermometer format tends to elicit strong response patterns that are at least partially independent of the stimulus content of the questions" (p.772).

While these findings are troubling in and of themselves, additional problems may arise in multivariate analyses that employ feeling thermometers. Scholars are rarely interested in the direct bivariate effects of feeling thermometers on opinions. Instead, thermometer scores are often used in combination with a host of standard demographic and political variables – the "usual suspects," such as education, income, party identification, and political values – to assess the background correlates of individual opinions. In a multivariate setting, interpersonal incomparability in thermometer responses could create additional problems if such incompatibilities are determined, at least in part, by the other variables used as controls in analysis. To demonstrate this problem, we follow Brady's (1985) formulation of the problem of interpersonal incomparability of responses. Assume that each individual *i* has a latent feeling towards a particular group or political figure represented by the variable T_i. We do not observe T_i. Instead, we observe the realization of the individuals mapping of their underlying feeling towards the political object onto the NES 101-point response scale. Like Brady (but using different notation) we assume that this mapping takes the form:

$$FT_i = NP_i + SF_i(T_i) \quad (1)$$

Where FT_i is individual *i*'s answer to the 101-point feeling thermometer question, NP_i is the individual's "neutral point"[17] – the point at which they feel neither positive nor negative towards the group relative to other groups evaluated – and SF_i is the individual's "scaling" factor – the degree to which they are extreme or moderate in their translation of their underlying evaluation to the NES response scale.[18]

To illustrate the problems caused by this transformation, imagine we are interested in the effect of the underlying evaluation of a political object, T on a dependent variable Y, controlling for another variable X. Because we observe only the transformation of T onto the NES response scale, we must instead use in our analysis FT. So we would estimate the equation:

$$Y_i = \beta_0 + \beta_1 FT_i + \beta_2 X_i + \varepsilon_i \quad (2)$$

Substituting in Equation 1, we get:

$$Y_i = \beta_0 + \beta_1 (NP_i + SF_i(T_i)) + \beta_2 X_i + \varepsilon_i \quad (3)$$

And factoring out β_1:

$$Y_i = \beta_0 + \beta_1 NP_i + \beta_1 SF_i(T_i) + \beta_2 X_i + \varepsilon_i \quad (4)$$

If NP and SF are constant ($NP_i = NP$ and $SF_i = SF$ for all i), meaning that the feeling thermometer scores are interpersonally comparable, then our analysis will proceed without difficulty. The B_1NP term will bias the constant, because NP does not vary across individuals. Since we rarely have substantive interest in the constant term this is usually not a problem. Similarly, if SF does not vary across individuals, it also poses little problem. With constant SF, it is simply a multiplier of the T variable – the SF term merely fixes the scale of the latent variable T. We will estimate B_1, which represents the effect of T on Y up to a scale factor. The "scale" of the latent T is irrelevant.

If, however, NP and SF vary across individuals, a host of problems may ensue. As Brady notes, the coefficient B_1 will be biased. In addition, in a multivariate setting the inclusion of control variables may

[17] This term is comparable to Brady's "origin" or α parameter.

[18] Another way to think of this is as an individual's variance in their use of the scale.

introduce a new set of problems. First, consider the case where NP_i is unrelated to the other variables, X, in the model. In this case, it acts like random measurement error, which biases *all* the coefficients in the model (Greene 1997, 435-444; Hanushek and Jackson 1977, chapter 10).

While this is troubling, it is nothing new in the world of survey data, where all measure presumably include some random measurement error. However, if either NP or SF have non-zero covariance with X, then there will be an additional source of bias in the coefficients B_1 and B_2. The shared covariance between X and the respondent's translation of their underlying political evaluations to a thermometer score (represented by NP_i and SF_i) will create what looks like covariance between X and the thermometer score, but is in fact, a measurement artifact due to the correlation between X and NP and SF. The coefficients on both FT and X will be biased. Here is a (hopefully) intuitive explanation:[19] assume that NP_i is systematically related to X, and we are regressing Y on FT and X. In this case, the variable FT is a combination of the true score T and of X. The coefficient we estimate for FT will, therefore, be a combination of the "true" coefficient (the effect of T on Y) and of the coefficient on X (the effect of X on Y). The coefficient on FT will inappropriately "soak up" some of the effect that should be attributed to X. In effect, the variance in the dependent variable that is properly explained by the independent variable X will be confounded by an artifact relating to how different individuals use what we presume to be the same thermometer scale.[20]

This result has potentially serious implications for how feeling thermometers are used in the study of political behavior. If the same factors that are used as independent variables in analysis also affect how people translate their underlying evaluations of groups and political figures into feeling thermometer scores, the coefficients on both the thermometer score variables and the other independent variables may be biased. Put another way, if the feeling thermometer scores contain response effect variance that is properly explained by the other independent variables in the model, then our inferences about the effects of the variables will be incorrect. So, for example, if individuals who subscribe to the principle of political equality are both more likely to support spending for the poor and have a higher Neutral Point, then a regression of desired levels of

[19] This draws on Hanushek and Jackson 1977, p. 288.

[20] This logic obviously extends to the multivariate case.

spending on the poor on the respondent's egalitarianism score and on the "poor people" feeling thermometer will yield biased coefficients on both the feeling thermometer and the equality variable.

Analysis

While response effects may pose a threat to our analyses, it is not possible to directly estimate the impact of such bias. We cannot, after all, measure T, NP, or SF. However, it is possible to gauge how much of a problem response effects pose by *indirectly* estimating the nature and severity of inter-personal incomparability in the use of the feeling thermometer. Specifically, because each individual answers a series of feeling thermometer items in each NES survey, it is possible to look at individual response patterns to the feeling thermometer items and get a sense of its relationship to other political variables. If we compare how different individuals evaluate the same set of political objects, we can determine if individuals vary systematically in the way they use the thermometer scales.

In order to do this, we chose a series of paired feeling thermometers from the 1992 NES: Democrats, Republicans, conservatives, liberals, big business, labor, homosexuals, and fundamentalists. We dealt with a set of paired thermometers – rather than the entire set included in the 1992 NES – in order to minimize possible bias in our calculation of each individual's mean score. Since the NES included more groups that have been traditionally associated with the Democratic party, including them all would have created a spurious relationship between our estimate of NP and party identification and/or ideology measures.[21] Because our calculations include evaluations of pairs of groups from opposite sides of the political spectrum, we can use the mean of the evaluations to serve as the estimate of the Neutral Point. Respondents who tend to be more generous in their evaluations of groups from both the left and the right – those with relatively high Neutral Points – will tend to have higher means. Conversely, individuals who tend

[21] The mean and standard deviation variables were calculated as the mean and standard deviation, respectively, across the eight groups. Respondents who rated fewer than half (i.e. two) of the groups were excluded. In addition, if a respondent rated one half of a pair but not the other (i.e. they rated liberals but not conservatives), that score was not included in the calculation. This allowed us to generated mean and standard deviation scores for 2,195 of 2,485 respondents (eleven percent of the cases are missing data). Analysis conducted using mean and variance scores for *all* respondents who answered at least one of the feeling thermometer items (n=2466; one percent of cases are missing data) yielded extremely similar results, suggesting that the pattern of missing data does not pose a threat to our analyses.

to take a dim view of all political groups will have low average feeling thermometer scores.[22] In a similar way, we can use the individual standard deviations across the evaluation pairs to estimate the respondents' Scaling Factor. Those respondents who tend to magnify differences between groups when translating their underlying evaluations to a feeling thermometer score will show higher variances than those who minimize their underlying differences in the evaluations of those groups.[23] Using the estimates of the mean and variance, then, we can see precisely how much inter-individual variation in the use of the feeling thermometers really exists.

In Figures Nine and Ten, we present the histograms of the individual mean and standard deviations. There is indeed a great deal of variation in how individuals approach the thermometer probes. The "mean mean," so to speak, is 52.4, but the individual means on the paired items vary greatly – from 6.3 to 88.1. Similarly, while the mean standard deviation is 20.5, it ranges from zero to 54.8.

While this individual variation in the use of the feeling thermometer is troubling, as discussed above, we are particularly concerned if such inter-individual heterogeneity is predicted by the same factors that are typically included in multivariate analysis. To gauge the severity of this problem, we modeled the individual mean and variances as a function of variables commonly used as controls in analyses of individual political behavior. Like Wilcox, Sigelman, and Cook (1989) – who follow a similar procedure – we are interested in the causal determinants of differences in the use of the feeling thermometer scale. But unlike those authors, we approach the modeling enterprise with an additional purpose in mind; namely the estimation of the severity of bias in normal analyses of political behavior. Thus, we predicted response variation as a function of the

[22] One might argue that this information concerning the respondent's general warmth or coldness towards groups in general is important information. Certainly, this may be true in some analyses. However, the coefficient on the feeling thermometer variable in Equation 2 is intended to measure the effect of differences in the evaluations of a particular group across people. To the extent that people are starting from different (neutral) points, as explained above, statistical analyses will confound such differences in neutral points with general variation in feeling thermometer scores. Put another way, if we simply use raw feeling thermometer scores, we may confuse relative differences in evaluations of particular groups with relative differences in what individuals consider to be neutral evaluations.

[23] Of course an individual's standard deviation is not a perfect measure of their Scaling Factor. A respondent who rates most groups at 50, but – for unexplained reasons – rates one group at 100 will show a higher variance than an individual who uses a range of values around 50 to evaluate political figures. Thus, the use of variance as a proxy for the Scaling Factor may confuse random error with differences in Scaling Factors. However, we believe that the standard deviation does serve as a rough proxy of the Scaling Factor.

"usual suspects" in analyses of political behavior namely: age, race, education, income, region of residence, marital status, union membership status, party identification, liberal/conservative identification, and subscription to the principles of equality, trust in government, racial resentment, moral conservatism, and limited government. In addition, we included measures of political information to see if respondents who differed in their levels of political engagement used the thermometer scores in different ways.

The results of these analyses are presented in Table One. A number of personal characteristics systematically affect how respondents approach the feeling thermometer scales. Turning first to the correlates of the mean, we find that, consistent with Wilcox et al. (1989), members of minority groups are warmer in general to the political groups than are whites. Similarly, men are consistently cooler in their assessments than women. Moving beyond those authors' work, we find that those individuals who identify with either major party are generally warmer than those individuals who are independent or claim no party identification. Likewise, people who belong to unions, have children, and are in the middle of the income scale are generally warmer towards the groups. All of these effects are relatively minor – ranging from one to several thermometer points across the range of the independent variables. Some of the largest effects on the mean, however, are determined by the political principles to which individuals subscribe. Those respondents who score highest on the equality scale are about four points higher than those who score lowest on that scale. And, not surprisingly given the objects of evaluation, those who trust government place their means over nine points higher on the scale than those who lack such trust.[24]

Turning next to the determinants of individual variance, a somewhat different set of predictors emerges.[25] Those individuals who score high on the trust in government measure are less variant in their

[24] One concern about this result is that the effect of effect of trust in government as measured in this model is an artifact of the explicitly political nature of four of the groups included in our mean measure – that those who trust government extend that positive feeling to the actors involved in government. However, the effect of trust in government is extremely robust across many measures of the mean, including ones made up of only groups that are quite explicitly *not* political. This might imply that the trust in government scale taps a dimension of general optimism or positivity – a matter beyond the scope of this paper.

[25] The use of regression analysis to model variance is not technically appropriate. Variances (and standard deviations) are bounded at the floor by zero. We are confident of our results however, because the OLS model did not generate any out-of-bound predictions, and because rerunning the analysis using a logarithmically transformed dependent variable yielded identical results.

answers, probably because of their higher mean scores. But the strongest set of determinants relate to the respondents' psychological connection with the political world. Respondents who are most politically connected – those who identify with a political party or with an ideological view – are more variant in their evaluations. This finding may not, however, be due only to differences in the Scale Factor between the politically engaged and the less engaged. Because the objects of evaluation here are political, it is probable that the politically engaged actually do hold stronger opinions – positive or negative – toward those groups. Conversely, it is likely that those who do not identify ideologically or with a party are less variant precisely because they feel less strongly about political objects, and rate them all closer together and toward the middle of the scale.[26]

The political value of moral traditionalism also has a strong positive effect on response variance: those who are most traditional have a standard deviation over six points less than those who are least traditional. Again, this is probably a substantive phenomenon. The moral traditionalism scale measures respondents' support for traditional norms, and opposition to new and relativistic moral codes. It seems probable that moral traditionalists see political and social groups in relatively black and white terms, or – perhaps more aptly – in hot and cold terms.[27] Finally, education has an effect on response variance. The combined effect of the education and squared education terms suggests that the least educated (less than eight years) are between two and four points less variant than other groups.[28]

These results may give pause to dismissing systematic differences in variance as a mere artifact of interpersonal incomparability. The largest effects – those associated with political engagement and with moral

[26] In contrast to the case with trust in government discussed above, the effect of the political engagement variables on response variance diminishes substantially when the dependent variable is a measure of the standard deviation across non-political groups. This lends further support to the idea that the political engagement effects are substantive and not due to scale factor differences.

[27] Part of the effect found here may be due, in part, to the strong negative reaction of the morally conservative to homosexuals. But even accounting for this fact, the morally conservative exhibit higher variances than those who do not subscribe to such values. The same analysis presented in Table 1, but with the homosexual and fundamentalist feeling thermometers removed from the construction of the dependent variable, yields a coefficient of 2.312 on the moral conservatism variable, which is significant at the 0.05 level.

[28] This probably is an effect due to the Scale Factor: the least educated simply use the extreme points on the scale more often than other respondents. Casual examination of the number of times respondents used the points 0, 50, and 100 seems to support this.

traditionalism – seem more likely to be a reflection of actual attitude variance than an artifact of individual use of the thermometer scale. Consistent with these findings, Poisson regression analysis indicates that the politically engaged use a greater number of distinct scale points than the less engaged. Together, these results suggest that – as is to be expected – the politically engaged have stronger and more varied views of the political groups they are asked to evaluate than do the less engaged.

The results relating to individual Neutral Points are somewhat more troubling. Although the effects are not huge, there is variation in individuals' neutral point on the thermometer scale, and that variation is related to political variables. Overall, we would expect an African-American female who is highly egalitarian and fully trusts the government to rate groups 18.9 points higher – almost a fifth of the scale – than a white male who is fully anti-egalitarian and scornful of government. This differences is especially significant because, as demonstrated by Equations 1-4, this systematic difference in neutral points across individuals may be confused with the true quantity of interest – namely, differences in how different individuals rate the same political group.[29]

IV. Evaluating Potential Solutions and Conclusions

People have proposed a number of solutions to concerns about the thermometer scale, most of them ad-hoc. Which makes the most sense obviously depends on the specific application. In this section we conclude our analyses by making some preliminary observations about the types of solutions available, and their advantages and costs.

Before beginning any analysis using thermometer scores, researchers should determine whether they think the thermometer is "interval-enough" for their needs. The most conservative approach would be to collapse the thermometer scores into discreet categories to use as rank-order data. So, for example, one could create dummy variables corresponding to "prefers the Democratic candidate" or "prefers liberals" and so on. This strategy is limited in two regards: first, it can only be used in situations where paired thermometer scores exist for comparison. That an individual "prefers illegal immigrants to Dan Quayle" might be an interesting

[29] This logic, of course, extends to analyses of political figure feeling thermometers.

fact, but it is not likely to be useful in most political analysis. Second, and perhaps more importantly, such an approach throws away a great deal of information. Our findings, after all, suggest that, for most people feeling thermometers convey at least as much – if not more – information as a 5- or 7-point scale. Thus, the use of feeling thermometers as a meaningful (semi) interval measure of respondents' evaluations of groups and individuals seems valid (at least for those thermometers examined in detail above

But our results do not suggest that thermometers should necessarily always be used "as is." Though feeling thermometers may approximate interval-level data, they are plagued by inter-personal incomparability. The effect of this incomparability on analysis may not be severe, but the results presented here suggests that researchers should consider addressing these measurement issues. Researchers may, therefore, want to correct the thermometer scales for the individual neutral point and/or scale factors differences. Given our findings, we would recommend the former, but not necessarily the latter. When it is possible, this is most easily done by simply subtracting two thermometers; use "black thermometer score" minus "white thermometer score" as measure of prejudice, for example rather than one or the other. But, as we just noted, creating such difference scores is often not possible. In this case, one might subtract off the individual mean – as estimated by averaging a series of bipolar pairs – from the raw thermometer score to purge the effects of inter-personal incomparability in neutral points. Beyond this, however, we do not suggest additional changes. While differences in the scale factor may vary across individuals in a systematic manner, accounting for this variation by, for example, normalizing the feeling thermometers by each person's individual mean and variance could lead a researcher to dismiss as artifact meaningful differences in the way that individuals evaluate political objects.

In sum, our work suggests that feeling thermometers are useful constructs. The thermometer scale, while not perfect, does convey significant information about respondents' affect toward political groups. The area of greatest concern we have uncovered relates to the Neutral Point, or mean thermometer ranking; it probably makes sense to correct this if possible. In future work, we plan to expand our analysis of the sources of random and non-random error in thermometer rankings, making use especially of over-time data to gain further leverage on the question of how thermometer scores should be used in analysis..

Table One
Models of the Mean and Standard Deviation Across Eight Paired T-Scores

Variable	Mean T-Score	Std. Deviation
Age	-0.095	0.013
	(0.080)	(0.075)
Age Squared	0.002	0.000
	(0.001)	(0.001)
Male	-1.056 ^	0.756
	(0.539)	(0.507)
Black	4.657 **	0.758
	(0.818)	(0.770)
Hispanic	4.376 **	-1.641 ^
	(0.920)	(0.866)
Education	0.849	-12.679 **
	(3.149)	(2.964)
Education Squared	-3.105	9.307 **
	(2.819)	(2.654)
Income <10k	-0.749	2.494 **
	(0.779)	(0.734)
Income 10k-20k	-2.157 **	1.649 **
	(0.675)	(0.636)
Income 40k-90k	-1.225 *	0.719
	(0.573)	(0.540)
Income 90k+	-2.462 *	-0.884
	(1.091)	(1.026)
Income n/a	-1.800 ^	2.487 **
	(0.970)	(0.914)
Deep South	0.853	2.466 *
	(1.034)	(0.974)
Perif. South	0.720	1.588 *
	(0.718)	(0.676)
Border States	1.778 ^	1.383
	(0.956)	(0.900)
Western States	-0.217	0.836
	(0.602)	(0.567)
Grew up in South	0.327	1.989 **
	(0.757)	(0.713)
Married	-0.574	-0.180
	(0.499)	(0.470)
Children	1.299 ^	-0.714
	(0.681)	(0.641)
Male X Children	-1.286	1.177
	(0.934)	(0.879)
Union Member	1.262 *	0.421
	(0.600)	(0.565)
	(continued)	(continued)

Table One, continued

Variable	Mean T-Score (continued)	Std. Deviation (continued)
Republican	1.583 **	2.829 **
	(0.598)	(0.563)
Democrat	1.491 **	2.031 **
	(0.545)	(0.514)
No Party ID	-2.345 *	-0.144
	(0.958)	(0.902)
Other Party	0.631	0.083
	(4.628)	(4.357)
Liberal	0.028	2.108 **
	(0.676)	(0.636)
Conservative	0.929	2.902 **
	(0.610)	(0.575)
No Ideology	0.657	0.174
	(0.651)	(0.613)
Egalitarianism	3.900 **	1.050
	(1.352)	(1.272)
Trust in Govt	9.327 **	-3.640 **
	(1.040)	(0.979)
Racial Resentment	0.258	1.562
	(2.044)	(1.925)
Moral Traditionalism	-1.182	6.076 **
	(1.143)	(1.076)
Limited Government	-1.864 **	-0.364
	(0.663)	(0.624)
Black Interviewer	-0.840	0.651
	(1.327)	(1.249)
Political Information	-2.770 *	1.496
	(1.307)	(1.231)
Constant	50.892 **	12.667 **
	(2.867)	(2.700)
n	1892	1892
R Squared	0.180	0.150
Std. Error of Estimate	9.160	8.620

Source: 1992 NES. Dependent Variables are mean and standard deviation, respectively, of respondent's thermometer scores of four pairs of objects: Democrats, Republicans, Liberals, Conservatives, Big Business, Labor, Homosexuals, and Fundamentalists. Thermometer scores underlying the dependent variables are coded zero to one hundred; all other variables are coded zero to one.

Cell entries are OLS regression coefficients, with standard errors in parentheses.
** $p<0.01$; * $p<0.05$; ^ $p<0.10$ two tailed

Figure One

Stylized Comparison between Thermometer Differences and Indicator Variables

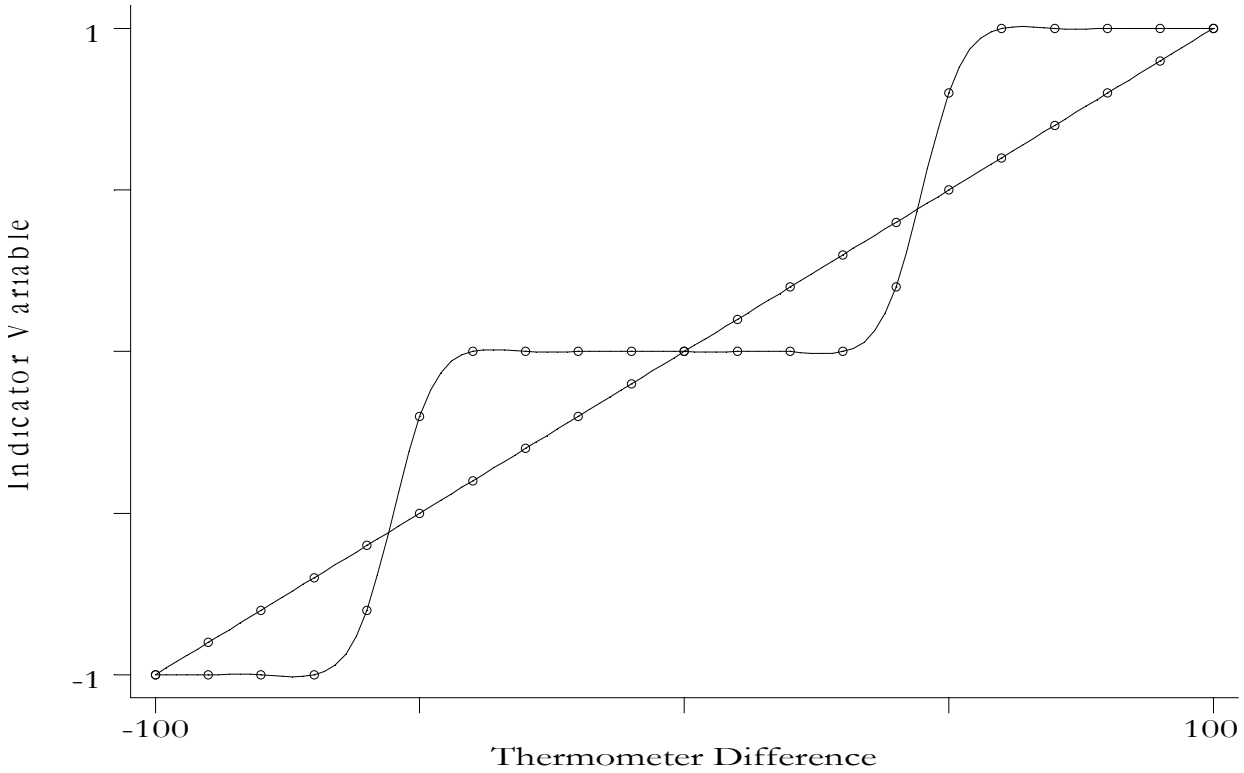
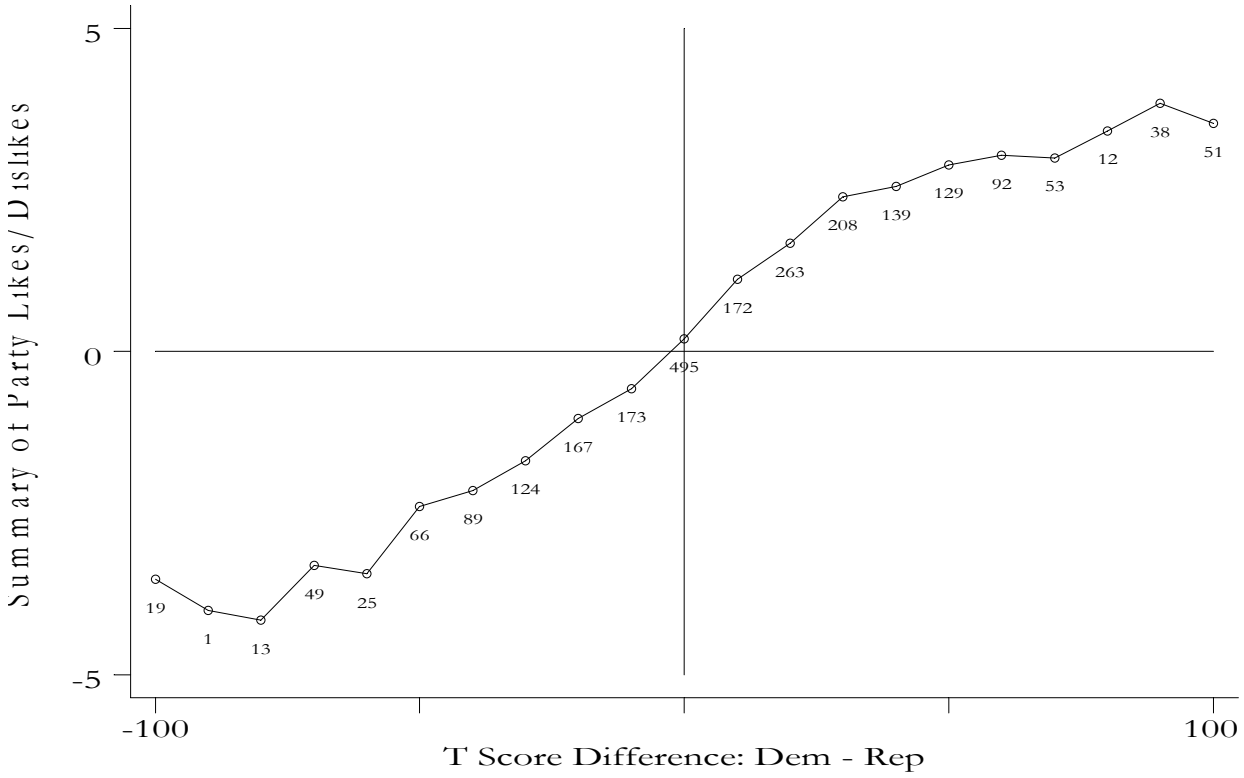


Figure Two

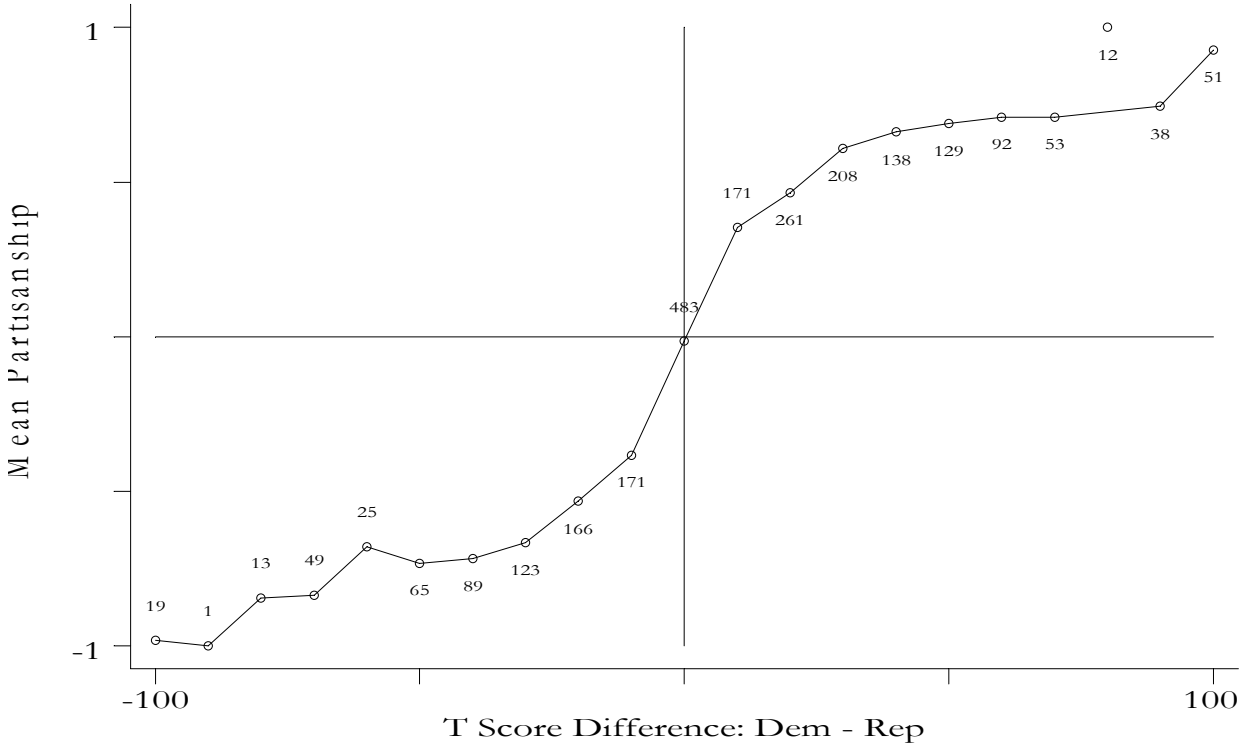
Mean Summary of Party Likes/Dislikes by Party Thermometer Difference



Source: 1992 NES. Numbers indicate number of cases used to calculate each mean.

Figure Three

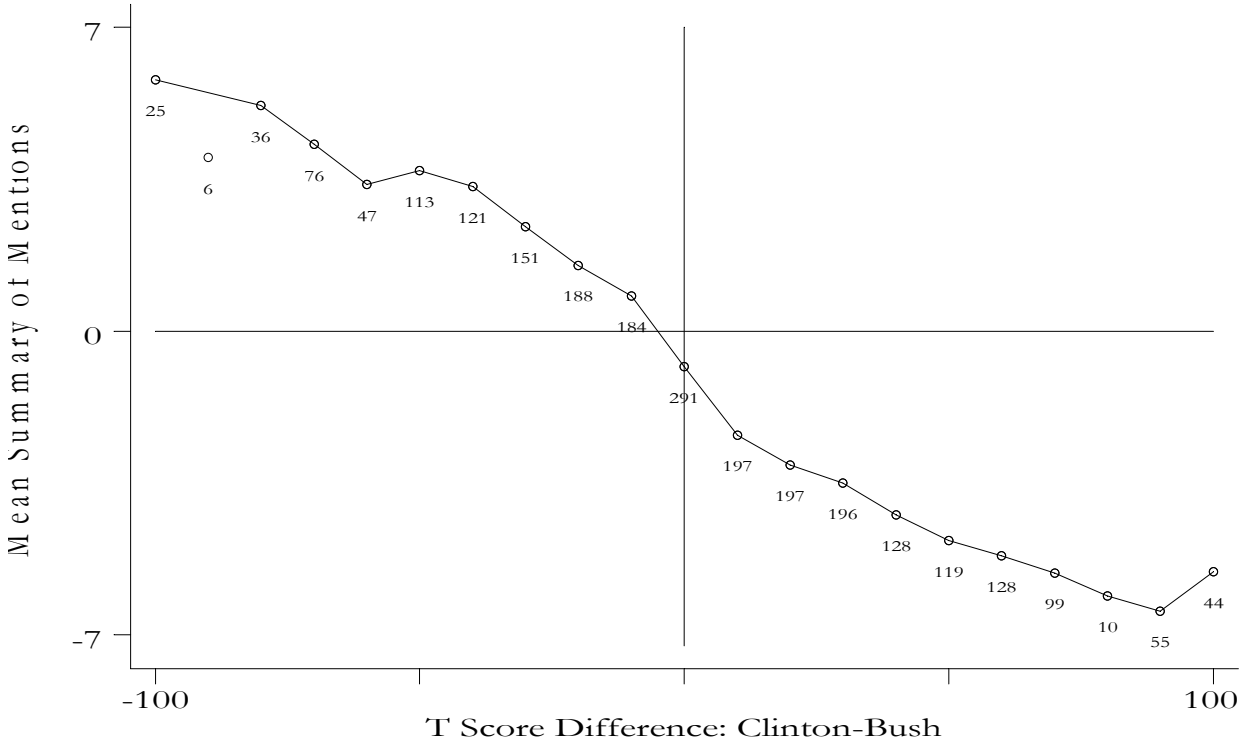
Mean Partisanship by Party Thermometer Difference



Source: 1992 NES. Numbers indicate number of cases used to calculate each mean.

Figure Four

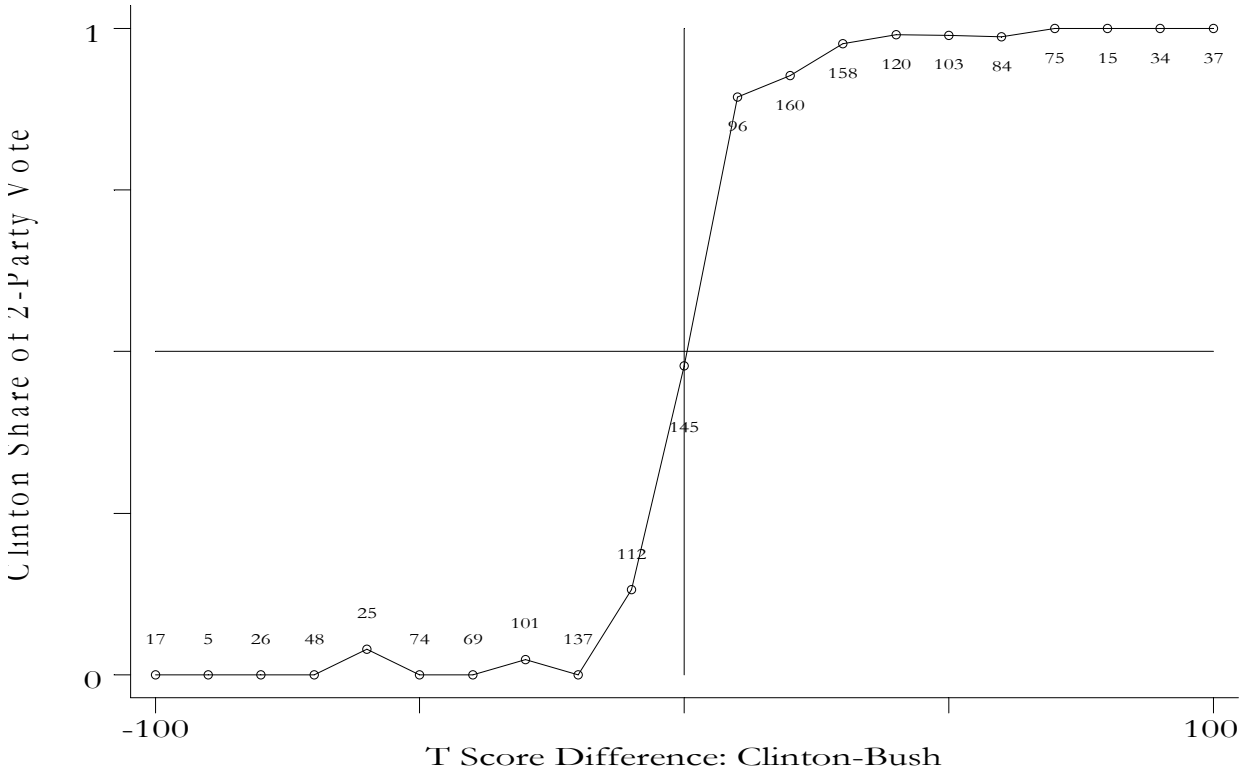
Mean Summary of Candidate Likes/Dislikes by Candidate Thermometer Difference



Source: 1992 NES, pre-election study. Numbers indicate number of cases used to calculate each mean.

Figure Five

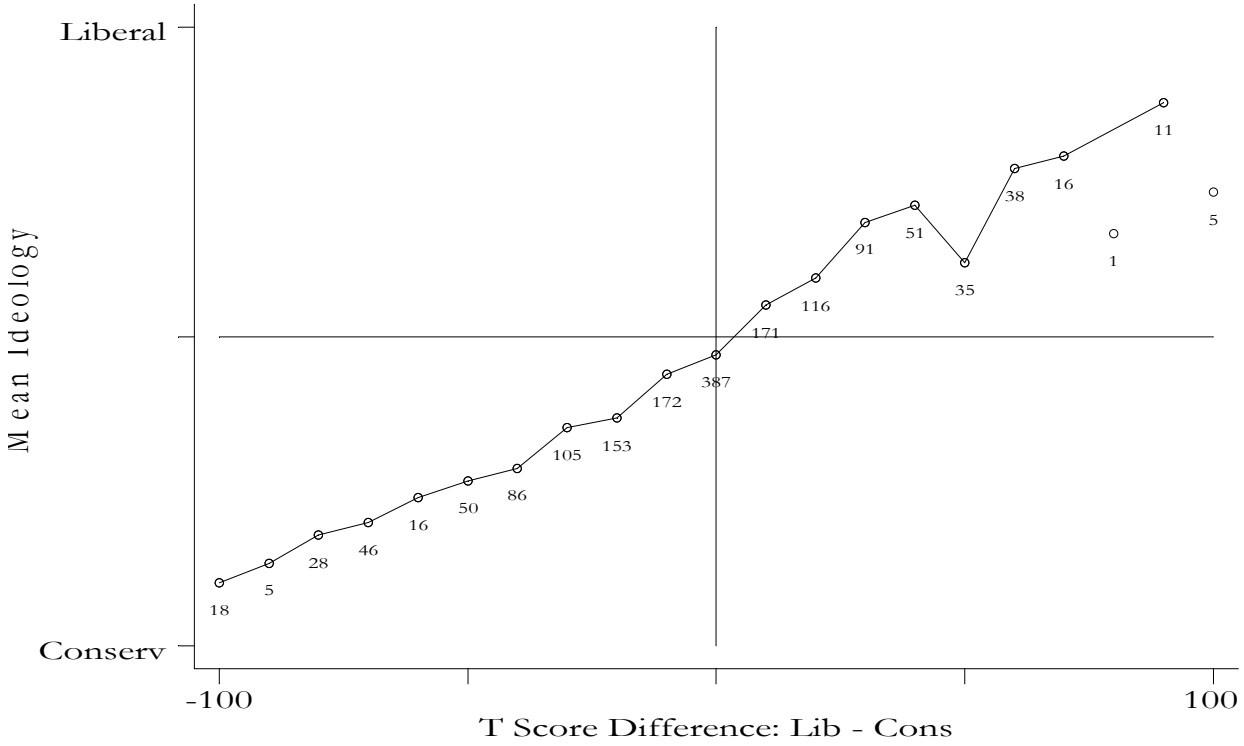
Mean Vote Choice by Candidate Thermometer Difference



Source: 1992 NES, post-election study. Numbers indicate number of cases used to calculate each mean.

Figure Six

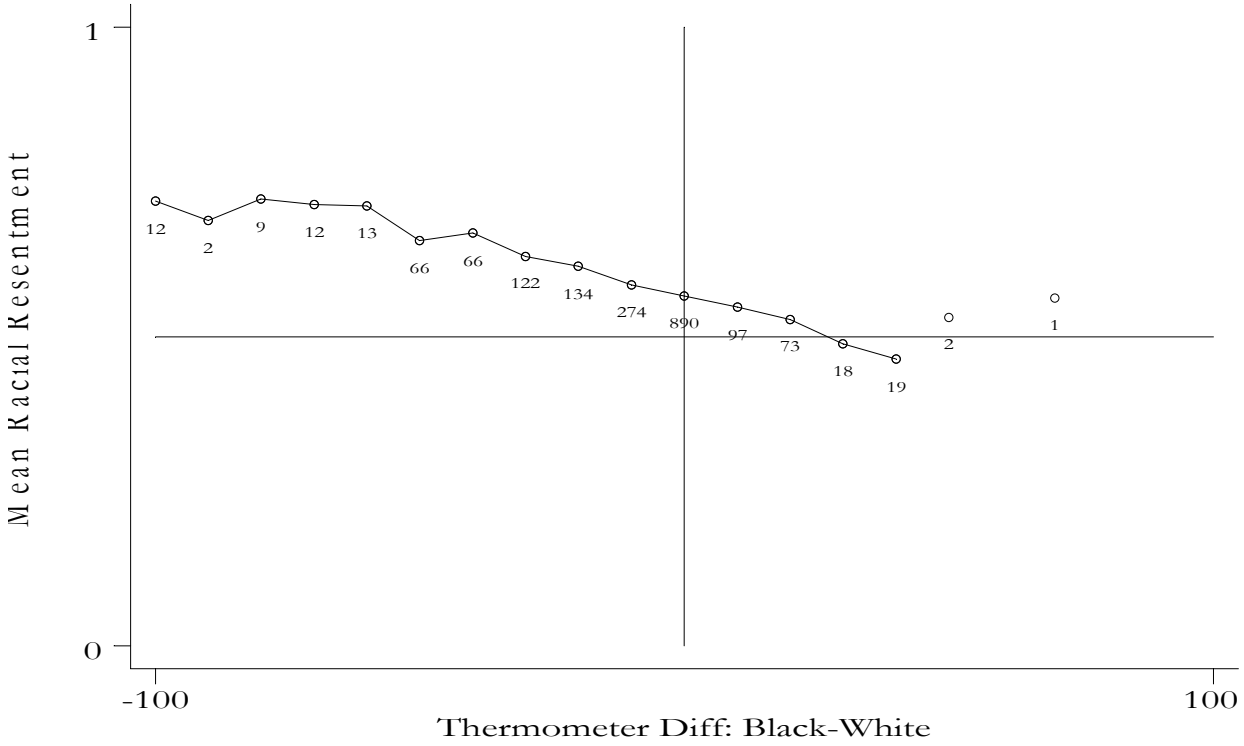
Mean Ideology by Thermometer Difference



Source: 1992 NES. Numbers indicate number of cases used to calculate each mean.

Figure Seven

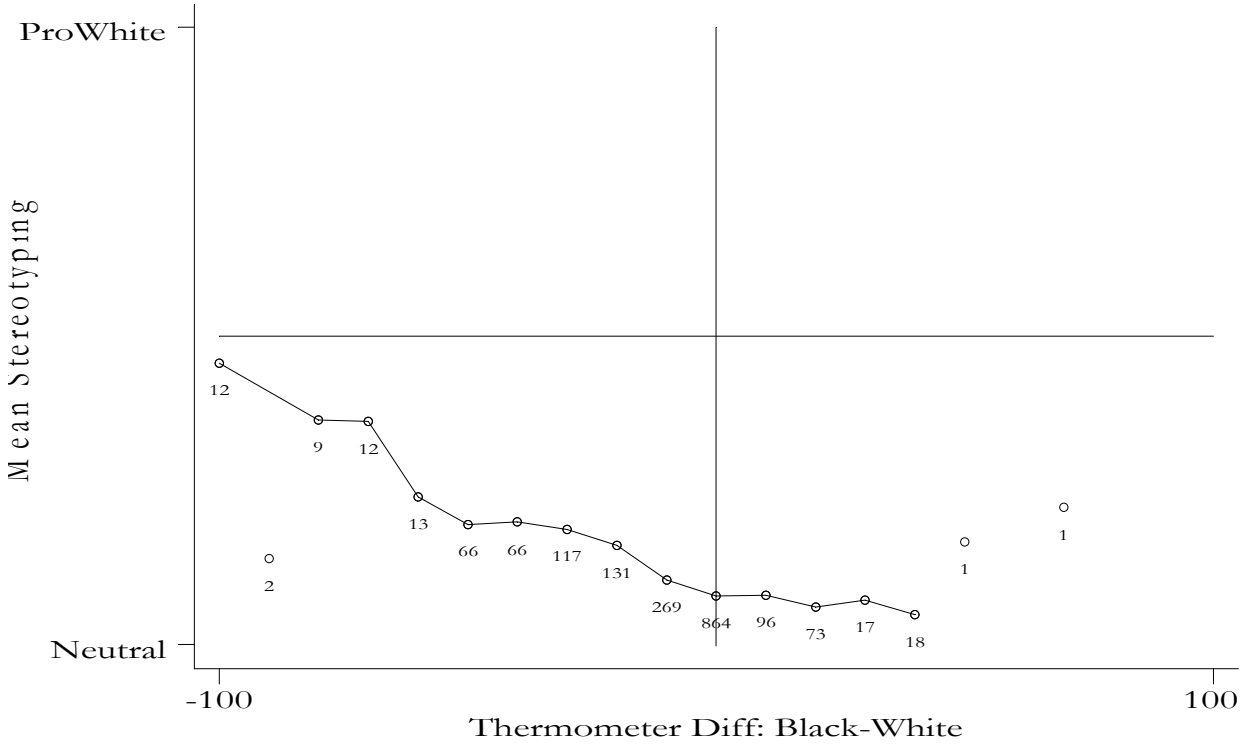
Mean Racial Resentment by Thermometer Difference



Source: 1992 NES. Numbers indicate number of cases used to calculate each mean.

Figure Eight

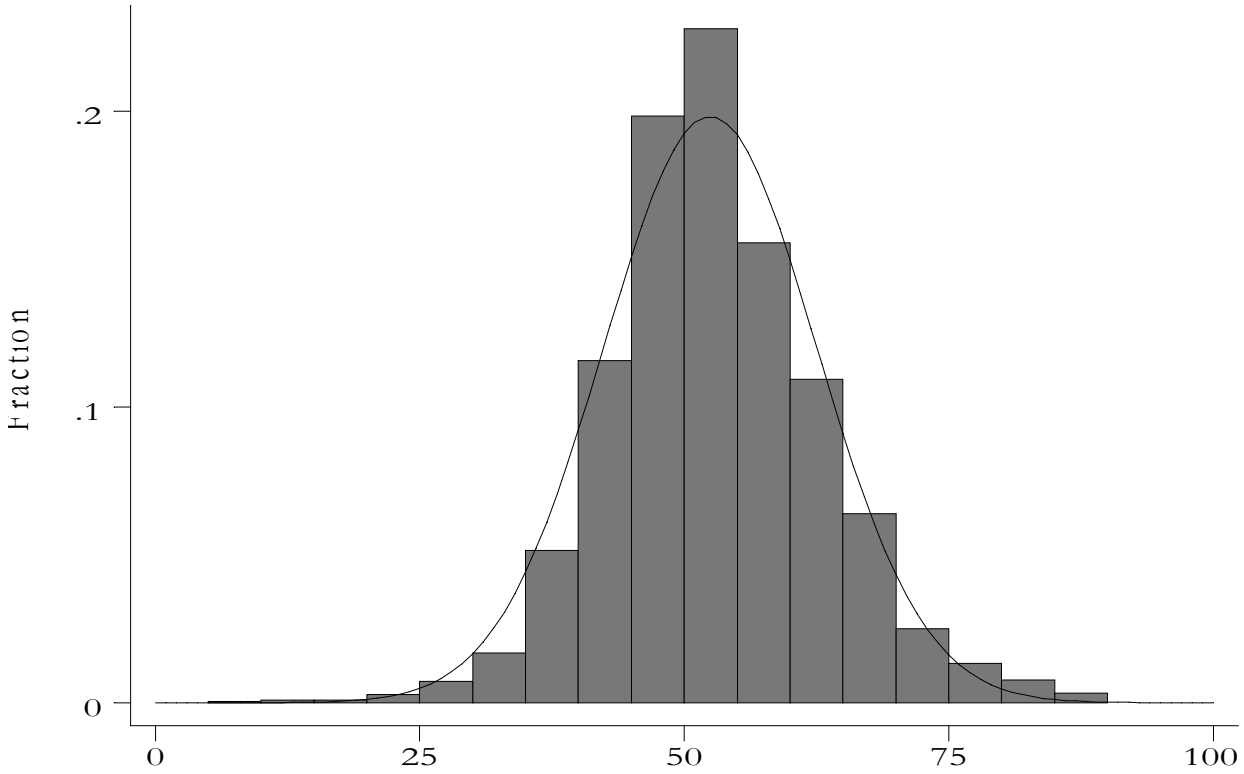
Mean Stereotyping by Thermometer Difference



Source: 1992 NES. Numbers indicate number of cases used to calculate each mean.

Figure Nine

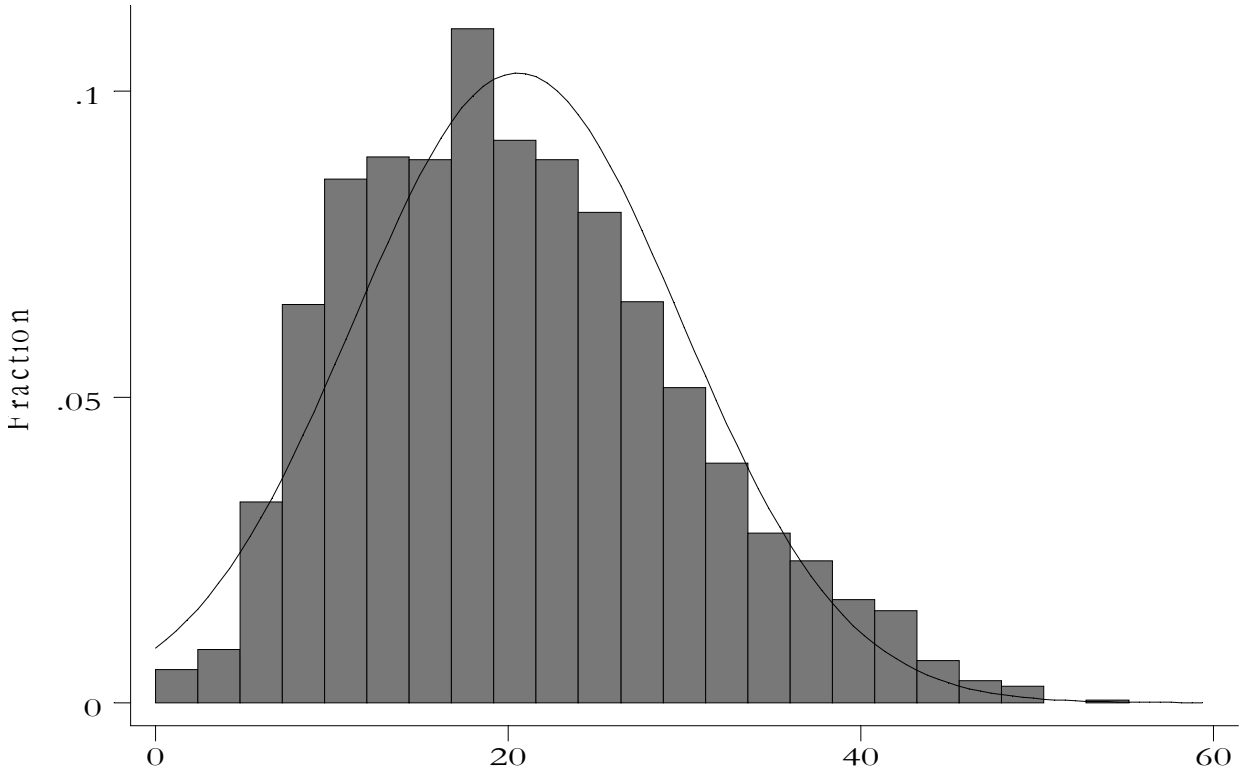
Mean Across Eight Paired Thermometers



Source: 1992 NES. Mean is calculated across four pairs of thermometers: Democrats, Republicans, liberals, conservatives, labor unions, big business, homosexuals, and fundamentalists.

Figure Ten

Standard Deviation Across Eight Paired Thermometers



Source: 1992 NES. Standard deviation is calculated across four pairs of thermometers: Democrats, Republicans, liberals, conservatives, labor unions, big business, homosexuals, and fundamentalists.

References

Brady, Henry E. 1985. "The Perils of Survey Research: Inter-Personally Incomparable Responses." *Political Methodology* 11:269-291

Green, Donald P. 1988. "On the Dimensionality of Public Sentiment toward Partisan and Ideological Groups." *American Journal of Political Science* 32:758-780

Greene, William H. 1997. *Econometric Analysis*. New York: Prentice Hall

Hanushek, Eric and John Jackson. 1977. Statistical Methods for Social Scientists. New York: Academic Press.

Wilcox, Clyde, Lee Sigelman, and Elizabeth Cook. 1989. "Some Like It Hot: Individual Differences in Responses to Group Feeling Thermometers." *Public Opinion Quarterly*. 53:246-257.

Zaller, John and Stanley Feldman. 1992. "A Simple Theory of the Survey Response: Answering Questions versus Revealing Preferences." *American Journal of Political Science* 36:579-616

Zaller, John. 1992. *The Nature and Origins of Mass Opinion*. New York: Cambridge University Press.