Title

Optimal Scale Length and Single-Item Attitude Measures: Evidence from Simulations and a Two-Wave Experiment

Authors

Goggin, Stephen
Stoker, Laura

Publication Date

2014

Peer reviewed

Optimal Scale Length and Single-Item Attitude Measures: Evidence from Simulations and a Two-Wave Experiment

Stephen N. Goggin
goggin@berkeley.edu
Travers Department of Political Science
University of California, Berkeley

Laura Stoker
stoker@socrates.berkeley.edu
Travers Department of Political Science
University of California, Berkeley

Abstract

Research on the optimal number of survey response options (scale length) has focused on multi-item indices, leaving us with less understanding of how the chosen length of a single-item attitude measure can affect its reliability, validity, and performance in statistical analyses. We employ both Monte Carlo simulations and a two-wave experiment to address these questions. The simulations provide an important basic framework for conceptualizing the issues at play in optimal scale length, while the experiment provides evidence against which the simulation-based expectations can be compared.

Paper prepared for the 2014 American Political Science Association Annual Conference, Washington DC

Introduction

What is the value of obtaining fine-grained vs. coarse measures of attitudes, opinions, preferences, experiences, and other subjective phenomena? A remarkably large number of studies from a multiplicity of fields have considered this question, with the earliest ones dating back at least to the 1920s. Every general overview of survey question-wording covers the issue (e.g., Krosnick and Presser 2010, Shaeffer and Presser 2003), and at least one author has taken up the challenge of providing a detailed review (Cox 1980). The topic is addressed under a variety of headings, including scale granularity (e.g. Brudvig 2007, Pearse 2011), coarseness, (e.g., Aguinis, Pierce and Culpepper 2008, Krieg 1999), sensitivity (e.g., Elkins 2000) and precision (e.g., Shively 1998). Even more common are writings that depict the question as involving the optimal length of a scale or the optimal number of scale points or response options.

Complicating matters is the fact that the scale length problem varies along multiple dimensions. How one thinks about or studies it depends, first, on the kind of property being measured. Is it a judgment about some stimulus that is external to the self (e.g., the usefulness of a given search engine) or about one's own feelings or beliefs? If it is about the self, is it, for example, about the extent of agreement or disagreement with some statement, the positivity/negativity of feelings about a political candidate, or the degree of importance placed upon some value? Another dimension is whether the measure will serve as a standalone indicator of a concept or instead as one of a set of component variables to be combined into an index. The survey mode is relevant, especially whether it allows visual displays to be available to respondents, as is the fact that changes in the number of response options usually coincides with changes in the number and meaning of response labels. Comparing odd-numbered to even-numbered scales (e.g., a 5-point scale vs. 4-point scale, or a scale ranging 0-10 vs. one ranging from 1-10) ties the scale-length question to the question of whether a middle alternative should be offered.

What is more, the question of how many scale points is optimal when designing survey questions is related to the question of what happens when one collapses a long or continuous scale into a smaller number of categories, which may or may not entail moving from a higher level of measurement (interval)

to a lower one (ordinal). As will become evident later, we see these questions as intimately connected, but the connection is often not made in the literatures addressing the twin topics, though exceptions to that generalization are still plentiful (e.g., Bollen and Barb 1981, Ferrando 2003). The literature on optimal scale length is peppered with evidence on how scale length influences reliability, while assessments of validity are scarce. The writings on collapsing otherwise continuous variables may touch on reliability but otherwise tend to hone in on problems related to bias and efficiency in statistical estimation and to consider non-survey data as well as survey data (Owen and Froman 2005 is a good example and also provides a nice overview of the literature).

We dip our toes into this water for three main reasons. First, despite the volumes that have been written on the topic of coarse vs. granular measures in survey research, very little of that attention appears to have focused on single-item indicators (Cox 1980), let alone single-item indicators of attitudes, which are our focus, here, and which are commonplace in the survey research of political scientists. It is not at all clear that one can draw lessons from the many studies of multi-item indices of Likert-type items when thinking about the optimal scale length of standalone attitude measures. Studies of scale length as it pertains to single-item attitude measures have mainly addressed the merits of using fully continuous graphic rating or visual analog scales (see, e.g., Couper et al. 2006 and references therein), though inquiries into web-based "sliders" have been trickling out (e.g., Cook et al. 2001).

Second, we feel that greater clarity can be brought to the topic by an approach that combines Monte Carlo simulations with an experimental study, which is what we present in this paper. Although there are plenty of simulation studies and empirical (some experimental) studies in the broader literature, it is rare to find a combination of the two. None that we are aware of focuses on single-item attitude measurement.

Finally, we seek to forge a tighter connection between the literature on scale length decisions in survey research and that which is focused on the statistical analysis of coarsened or collapsed measures. This means going beyond the focus on how the reliability of an attitude scale varies depending on whether it is measured in 5 vs. 7 vs. 11 vs. 101 points, to consider how the same measurement decision

affects the ability to reach valid conclusions when the measure is put to use as a dependent or as an independent variable in a statistical analysis.

In what follows we begin by laying out the logic and design of a set of simulations that we use to demonstrate how the properties of an attitude measure may vary as a function of the length of its scale. The simulations reflect a model of the survey response, which we articulate, and build in variations to reflect matters of debate or uncertainty, which we describe. A central feature of the approach we take is found in its delineation of random error vs. rounding error in measurement. The simulations demonstrate how reliability, validity, and effect estimation can vary as a function of the length of the attitude scale. After presenting the results of the simulations and drawing lessons from them, we report on results from an experimental study designed to be complementary.

Simulation Logic and Design

To begin, we assume that each individual's attitude can be represented as a distribution over a real-numbered attitude scale, which is set to range from 0 to 100.[1] The center of that distribution is the person's *true score*, while the variation indicates how a person's responses are likely to fluctuate across repeated measurements. The *actual response*, the response given on a single occasion, is merely a point from the distribution. The individual then faces a mapping problem, of converting his or her actual response into a position on the scale that is presented by the survey question. We refer to the choice made as the individual's *observed response* (or observed score). Of paramount interest is how well observed responses reflect true scores as the attitude scale presented to respondents varies in length (i.e., the number of scale points or response options).

As emphasized by Achen (1975) and many scholars writing subsequently (e.g., Alwin and Krosnick 1991, Green and Palmquist 1994), fluctuating responses can be thought of as being induced by the psychology of the individual respondents or by the measurement process/instrument itself. According to one version of the individual instability argument, people's responses will vary depending upon the

[1] Later, we replace the assumption that true scores can take real numbered values between some limits with an assumption that true scores can take integer values between some limits.

particular set of considerations that are salient when they are called upon to express their attitude (Zaller 1992). Variation in responses around a central tendency reflects the variation in considerations that the individual samples when he or she provides a response. A second version of the individual instability argument holds that people may not be able to reliably place themselves on a single point of the continuum but that they will be able to identify a range within which their opinion lies (Sherif and Hoveland 1961, Krosnick and Presser 2010). If asked to select a specific point on the continuum, a person would choose a point within that range. Finally, fluctuation in responses could be errors induced by vagueness or complexity of the measurement instrument or features of the setting in which it is implemented. We refer to these three ideas about the source of response fluctuations using the labels of *response sampling*, *judgmental uncertainty*, and *measurement error*, respectively.

Regardless of why responses fluctuate, we assume it is the central tendency of the distribution—the true score—that we are trying to measure. This assumption is warranted if it is the true scores that are causing and being caused by other variables of interest, not the response that deviates from the true score due to any given moment's sampling of considerations, or uncertainty about where, precisely, to place oneself within some range of the continuum, or errors of measurement. In addition, we assume that the variation within individuals across repeated measurements is random—that people randomly draw a score from their own response distribution, which is centered on their true attitude. This assumption is warranted if the sampling of considerations is as-if random (response sampling model), if respondents are certain of the range within which their opinion falls but uncertain as to where within that range it falls (judgmental uncertainty model), and/or if the sources of measurement error are unsystematic or idiosyncratic.[2]

To represent these different accounts of response fluctuation in our simulations we vary the form of the error distribution, using both a normal and a uniform form. The idea that responses deviate from true scores due to response sampling or measurement error—or some of both—is probably best captured

[2] We also assume that the errors are independent across individuals within the sample.

by the normal form, while the idea that fluctuations reflect judgmental uncertainty is best represented by the uniform. We also vary the standard deviation of the error distributions.[3]

This set-up allows the magnitude of random variation in people's responses to vary across simulations but not across individuals within simulations (with one exception, described below). Moreover, it assumes the extent of random variation in responses is the same regardless of the scale with which attitudes are measured. This, however, may be implausible, as it is commonly believed that longer scales are more difficult for respondents and may introduce more measurement error as a result. Thus, we also introduce a variation that allows measurement error to grow with scale length.

The first step in each simulation was the generation of true scores. We used two symmetric true score distributions: (a) normal, mean 50 and standard deviation 20; (b) uniform, mean 50 and ranging from 0 to 100.

Next, the actual response for each person was generated by drawing an error from the appropriate distribution (normal or uniform, with varying magnitudes of error variance) and adding that to the true score. This procedure yields some cases with responses less than 0 or greater than 100. In some simulations, these outlying scores were recoded to 0 and 100, respectively, which effectively forced the actual scores to fall between 0 and 100, inclusive. This censoring procedure reduces the error variation of individuals with true scores at or near the extremes of the scale, which captures the well-known tendency of people with more intense attitudes to show more response stability (Krosnick et al. 1993, Visser and Krosnick 1998).

The third step in the simulation was the mapping of actual responses onto response options, for scales of varying length. We represented this choice for scales with 2, 3, 4, 5, 6, 7, 11, and 101 points. The 2-point scale presents people with options for identifying the direction of their attitude, favorable or unfavorable. The 3-point scale adds a middle or neutral option. The 4 and 5-point scales allow for intensity variations (e.g., very favorable vs. somewhat favorable) but differ in whether a middle point is

[3] With normally distributed errors, we used standard deviations of 0, 5, 10, 15, 20, and 25. The uniformly distributed errors were set to range from +/- 0, +/-5, +/-10, +/-15, +/-20, and +/-25.

provided. The 6 and 7-point scales allow for further distinctions in intensity, while again varying whether the middle option is offered. Finally, the 11 and 101-point scales keep the neutral option but allow for finer and much finer intensity variations, respectively. Although other scale lengths are found in survey research (most notably, also 9, 10, and 21), these seem to be the most common. Our experiments use the same scale length variations.

In mapping their actual responses to scale points, we assume that people make choices following the proximity logic that underlies most spatial models: they choose the point on the scale that is closest to their own.[4] Our procedure assumes that a scale offering k points breaks the 0-100 continuum into k ranges, each of which is bisected by a midpoint that equals a response option on the scale. Thus, for example, a 2-point scale breaks the continuum into two ranges: 0-50, and 51-100, with midpoints of 25 and 75, respectively. If asked to respond on a 2-point scale, anybody with a response less than 50 would choose the first option and receive an observed score of 25, while anyone with a response greater than 50 would choose the second option and be scored as 75.[5]

A plausible alternative to this scoring of the response options is one that anchors the options to the end-points of the scale but otherwise divides the scale into equal intervals. This would yield scores of 0 and 100 for the 2-point scale, 0, 50, 100 for the 3-point scale, and so on up to the 101-point scale, which would be scored using 101 integers. Fortunately, the choice of one or another of these scoring methods is irrelevant to many results concerning the properties of the scales, as the two scoring methods are linear transformations of one another. The choice does, however, matter to some calculated quantities, such as the observed error variance of the scale, which should be kept in mind.[6]

---

[4] This mapping rule ignores the possibility of response sets, such as a systematic tendency toward picking less (or more) extreme options, or reporting a positive attitude when one's true position is neutral, or simply gravitating toward a given point on a scale (e.g., the mid-point), perhaps over and over in a battery (Krosnick and Presser 2010, Podsakoff et al. 2003, Pasek and Krosnick 2010). We do consider response sets in the experimental analysis.

[5] The full set of observed scores are as follows (but multiplied by 100): 2-point scale: 1/4, 3/4; 3-point scale: 1/6, 3/6, 5/6; 4-point scale: 1/8, 3/8, 5/8, 7/8; 5-point scale: 1/10, 3/10, 5/10, 7/10, 9/10; 7-point scale: 1/14, 3/14, 5/14, 7/14, 9/14, 11/14, 13/14; 11-point scale: 1/22, 3/22, 5/22, 7/22, 9/22, 11/22, 13/22, 15/22, 17/22, 19/22; 101-point scale: 1/202, 3/202, …, 101/202, …, 199/202, 201/202.

[6] As is evident, we assume that all scales are measured at the interval level, which accords with our sense of what most analysts do. Some researchers have addressed the scale length question by considering coarse measures as yielding ordinal data compared to the interval data of an uncoarsened scale. Cicchetti, Shoinralter and Tyrer (1985)

The key virtue of the approach taken in these simulations is that it distinguishes two ways that a discrepancy between a true score and an observed score could arise: (1) because response sampling, judgmental uncertainty, and/or measurement error lead an actual response to deviate from the true score, and (2) because actual responses have to be mapped onto the scale options supplied. This allows us to explore consequences of each of these errors, separately and together. We will refer to the first as random error, and to the second as rounding error.

If we assume no error of the first type, such that true-scores=actual scores, the differences obtained across scales are entirely due to rounding error. If so, the consequences of measuring a variable in a coarser manner than the underlying variable are equivalent to the consequences of collapsing a well-measured continuous variable. This brings two problems that are rarely considered together—collapsing vs. measuring coarsely in the first place—into the same conversation. Estimation biases and power limitations that come from the collapsing of continuous measures (see, e.g., Owen and Froman 2005, Shively 1998, ch. 5) are the same as those that can arise due to coarse measurement in the first place (see, e.g., Brudvig 2007, Krieg 1999).

Since the more reasonable expectation is that actual responses will, in fact, deviate from true scores, the question that then arises is whether rounding error can actually end up improving our measurement and analysis. The idea that rounding could help can be illustrated with a simple example. Suppose an individual's true score is 70 and her observed score (adding a random error) is 75. If she is asked to respond on a 101 point scale she will respond with 75, yielding an observed score that deviates by 5 points from her true score. If she is asked to respond on a 2-point scale, she will also respond 75, with the same error. But if she has the option of a 5-point scale, she will choose 70 (over 50 or 90), resulting in an overall error of 0. The rounding error, in this case, erases the random error. Examples like this raise the possibility that coarse scales may not be problematic even if true scores vary continuously.

do so when simulating the choice of a scale length, while Srinivasan and Basu (1989) do so when considering the consequences of collapsing a continuous measure.

We begin by considering what the simulations reveal about the reliability of variously coarsened scales. Following classical test theory, an error in measurement is the difference between the observed score and the true score. In our simulations, this error is the sum of the random error component and the rounding error component. Reliability, in turn, is the squared correlation of the true and observed scores. If there is no random error in a measure, only rounding error, coarsened scales are doomed to be less reliable than uncoarsened (or less coarsened) scales due to the fact that there is unmeasured heterogeneity in true scores among those obtaining the same observed score.[7] Whereas reliability would be perfect for the uncoarsened scale, it would drop as the number of scale points diminishes and the rounding error grows. Our simulations will illustrate this case.

The more interesting question is how reliability levels change across variously coarsened scales as random error is also entered into the equation. Here, it is useful to remember that reliability is not only equal to the squared correlation of the true and observed scores, but is also equal to the ratio of two variances:

Reliability = variance of true scores/variance of [true scores plus error]

On the one hand, rounding error will reduce the true variation in coarsened relative to uncoarsened (or less coarsened) scales. On the other hand, the fact that random errors can be diminished by rounding means that coarsened scales may not suffer as great a reliability loss as uncoarsened scales once random errors are present; i.e., a coarsened scale's smaller true variance may be more than compensated by its smaller error variance. We evaluate whether under the conditions of our simulations we are ever better off reliability-wise when opting for coarsened scales.

We also demonstrate that the standard test-retest strategy of assessing reliability can be misleading when coarsened scales are used. In the hypothetical world in which no error is random and all error comes from rounding, test-retest reliabilities would always equal 1. That is because a deterministic rule is leading directly from a true score to an observed score (the rule dictated by our spatial proximity

[7] Reliability is also equal to the eta-squared from a one-way analysis of variance with the true scores serving as Y and the coarsened scale scores serving as X.

logic). If the process of measurement were repeated, the same true and observed scores would show up again and again. But, of course, those observed scores are rounding error-ridden, and the true reliability of the measure is less than 1. We show that test-retest correlations based on coarsened measures will tend to overestimate the true reliability even when random error is introduced, though the extent of the bias dwindles as the random error variance increases.[8]

After presenting the reliability results of the core simulations, we turn to two important variations. Each variation speaks to an important idea in the literature on coarsened measures. The first allows for the possibility that random errors in response become larger in magnitude as the number of scale points increases. This reflects the idea that longer scales are more difficult scales for survey respondents to use, which has the consequence of introducing more noise into our measurement (Krosnick and Fabrigar 1997, esp. pp. 144-145, Krosnick and Presser 2010, esp. pp. 269-271). The second variation allows for the possibility that true scores fall only along a select number of discrete points along the 0-100 real number line; i.e., that attitudes are fundamentally discrete, not continuous, variables. This is an idea with a long heritage within the survey research field, fueled by early work suggesting that people could only differentiate about 7 degrees of difference when characterizing external stimuli, if not themselves (Miller 1956), and the regular finding that people do not use all of the points available on very fine-grained scales (e.g., Ferrando 2003; for further ideas and references along these lines, see Krosnick and Presser 2010, Schaeffer and Presser 2003). This variation of the simulation allows us to represent the rather obvious problem of what happens if we present respondents with more scale points than they are capable of discriminating.[9]

[8] We have not found anyone making this point in the literature on scale length. It is, however, well known that test-retest correlations are biased if the errors are correlated (e.g., Blok and Saris, 1984), and correlated errors are responsible for the upward bias in test-retest correlations when using coarsened measures, given our simulation set-up. Scholars have also shown via simulations and proofs that, certain assumptions given, the correlation of two different coarsened variables will be biased toward zero (e.g., Bollen and Barb 1981, Krieg 1999).

[9] Notice that we have already represented this idea in a different fashion, specifying that people can identify an interval on a continuum in which their true opinion lies but not a precise point. Unlike the discrete model, this still allows for a continuous range of true values.

After using the simulations to illustrate how the coarseness of a scale affects its reliability we turn to how coarseness affects a measure's validity. Here, the focus is on the extent to which coarsened scales contain systematic errors in measurement—errors that, in expectation, are non-zero—and what that entails for the estimation of causal effects. We first use the simulations to illustrate the properties of rounding error as it varies by scale length, demonstrating how rounding introduces systematic errors in measurement that are correlated with true scores. We then show how this validity problem with coarsened scales biases estimation of causal effects. In the interests of space, we focus on how analysts will get misleading answers when estimating how an experimental X affects attitudes measured too coarsely. However, the reliability and validity problems with coarsened scales will also confound inferences when estimating the effects of attitudinal Xs on other dependent variables.

Simulation Results: Reliability

Figures 1a-1d report two sets of results from the four sets of simulations.[10] Each simulation was run with a sample size of 1,000 observations, and was replicated 1,000 times. Averages across the 1,000 replications are presented in the figures. The top panel depicts true reliabilities (the squared correlation of true scores with observed scores) as well as test-retest reliabilities (the squared correlation of two sets of observed scores obtained under conditions of repeated measurement). The bottom panels depict error variances, where an error is defined as the difference between the true score and the observed score. As discussed earlier, the magnitude of this error for any one case in any given simulation is the sum of random error and rounding error.

A first finding of significance is that reliability is always a monotonically increasing function of the length of the scale. In no case does reliability drop as the number of response options grows.[11] This is so even when the error variance of the most fine-grained scale—the 101 point scale—comes to exceed

[10] True scores are normal or uniform, each with normal or uniform random errors, as described earlier.

[11] It is worth pointing out, however, that this is not what one finds unless one censors the data. If the simulations are run without censoring at 0 and 100, reliability tends to decline as one moves from an 11-point scale to a 101-point scale if there is a large amount of random error in the measurement. In this case the reliability dips when moving from the 11-point to the 101-point scale because the greater variation in true scores is more than offset by the greater error variation.

that of the coarser 11-point, as it does in each simulation that models the random error as coming from normal distributions with large variances (SDs of 20 or 25). For example, in Figure 1a, the error variance for the 101-point scale (515) exceeds that evident for all of the other scales except the 2-point scale when the random error variance is large (SD=25). As discussed earlier, this occurs because the 101-point scale does not benefit from rounding error to the extent that coarsened scales do, which can (and in this case, does) work to reduce the error variance in the coarser measures overall. Still, the reliability of the 101-point scale never suffers in comparison to the coarser scale because the greater error variance is more than compensated for by the greater true variance that is captured by the more fine-grained measure.

To say that the reliability never dips down when moving to a longer scale is not to say that the reliability gains from more fine-grained measures are always of great significance. Several finding in this respect are worth noting. First, the smaller the quantity of random error in the data the more one gains from using fine-grained measures. Thus, for example, differences in the reliability of the 5-point and 101-point scale are more substantial when the error SD is 5 or 10 in figures 1a and 1c than when it is double that or more. Put in terms of our earlier discussion about the sources of random error in measures, the implication is that fine-grained measures are more desirable when assessing attitudes that individuals hold with more certainty or that are sufficiently crystallized so as to show little variation as a result of the sampling of different considerations. Fine-grained measures would also be especially valuable if used in settings where measurement error can be minimized by, say, reducing respondent anxiety, fatigue, and distraction. While not surprising, the idea that fine-grained measures are especially valuable for some attitudes and if obtained in some measurement contexts is important and yet absent from much of the scholarly conversation about scale length.[12]

[12] That point should not be overstated, however, as the most fine-grained measure consistently outperforms the others when the random error is modeled as uniform (Figures 1b and 1d), regardless of the magnitude of the error variance. This is because the errors under the uniform assumption are strictly bounded, and cannot take on the extreme values that are possible under the assumption that errors follow a normal. Observations with extreme errors are high leverage observations where reliability is concerned.

Second, by far the poorest reliability showings are found for the 2-point and 3-point scales (especially the 2-point scale) in each of the simulations. The gap in reliability between the 2-point scale and those of middling coarseness (e.g., 5 or 7-point) is always greater than that found when comparing the middle-length scales to the 101-point scale, usually by a significant extent. This reflects the non-linear way that the true variance captured by the scale grows as the number of categories grows, a finding well-known from information theory (see, e.g., the discussion in Alwin 1992).[13] The power of this feature to shape reliability levels is evident from the consistency of this pattern across all of the simulations we carried out, regardless of how the other parameters were specified.

Third, it is useful to compare the scales in terms of how sensitive they are to the quantity of random error in the data. Not surprisingly, the reliability of the 2-point scale is the least responsive to variation in random error. Intuitively put, this is because much of that error, while producing variation in actual responses (as defined earlier), is leaving observed scores unchanged. Whether the magnitude of random error is small or large, the fluctuation in people's actual responses is likely to be within regions associated with a particular category (e.g., "favorable") rather than across them. As the number of response options increases, random variation in responses becomes more likely to be translated into observed variation in responses. Visually, one sees this in the fact that the reliability levels are more spread out on the right hand side than on the left hand side of Figures 1a-1d.

Finally, the consequences for reliability of varying a scale's length is at least matched and usually dwarfed by the consequences of variation in random error, at least as we have represented it in these simulations. If that error is induced by fuzzy-thinking or consideration-sampling respondents, the solution is not to resort to coarser measures, even if doing so may make matters only a little bit worse. Likewise, collapsing a continuous measure is not helpful. If some of that response instability is, instead under one's control—i.e., if it is measurement error, as we have described it—then it is that which should be the focus of one's efforts. Although the options here could include innovative ways of assessing

[13] The standard formula is that the scale's information carrying capacity equals $\log_2$(number of categories).

attitudes or controlling the context of measurement to minimize measurement error in single indicators, the most obvious way forward is through the construction of indices that combine parallel measures of the same construct to yield more reliable indicators.

Two other results from Figures 1a-1d are worthy of note, each of which we mentioned when discussing the logic of the simulation set-up. First, notice that in each of the figures reliability grows with scale length (top panel) and error decreases with scale length (bottom panel) even when there is no random error in the data (SD=0). Although this situation is purely hypothetical, it clearly reveals one cost of collapsing an otherwise well-measured, continuous measure, which Shively (1998, p. 60) aptly described as "the sin of wasting information."

Another important finding concerns the comparison of test-retest reliabilities with true reliabilities (top panels of Figures 1a-1d). As anticipated, test-retest reliabilities overstate the reliability of coarsened measures, by definition when there is no random error and only rounding error in the variables, but also or especially when the random error in the data is small in magnitude. In these cases, stability in the observed scores across repeated measurement is masking instability in actual (unobserved) scores within the regions associated with any given category. Interestingly, this bias tends to be negligible after reaching 5 categories and to diminish sharply as random error grows. This finding may be of particular relevance to debates regarding the measurement of democracy, despite being generated by simulations designed to speak to the measurement of attitudes. The value of dichotomous vs. more fine-grained measures of democracy is a vibrant topic of debate, in which reliability assessments play no small role (see, e.g., Elkins 2000).

Figures 2 and 3 report on simulations that, in turn, vary a key assumption of our first set. Whereas the earlier simulations assumed that any errors in measurement were constant across scales of varying length, those yielding the results in Figure 2 assume that the random error in the 101-point measure is greater than that of the coarser (2 through 11-point) scales. As discussed earlier, many discussions of scale length raise the possibility that longer scales might be noisier scales because of the complexity of the task that they present to respondents: discerning which of 101 points represents their attitude. Indirect

evidence in support of this idea includes the finding that respondents rate the 101-point scale lower than coarser scales in terms of ease of use and higher in terms of the time they needed to provide a response (though also higher in terms of its ability to allow them to express their feelings, Preston and Coleman 2000). More generally, it is common sense to expect different measurement instruments to vary in the measurement error they induce. The *Parts Express 390-722 Mini Digital Sound Level Meter* (which Google prices at $22) probably yields noisier observations than the *Extech 407790 Real Time Octave Band Analyzer Decibel Sound Meter* (which Google prices at $3,799).

Our core simulations also built in the assumption that true scores vary continuously along a scale bounded at 0 and 100. What if, however, people are simply not *capable* of making such fine-grained distinctions? As Krosnick and Presser (2010), among others, have pointed out, it is not at all clear that people know what to make of differences across a 101-point scale: "[O]nce the number of scale points increases above seven, point meanings may become considerably less clear. For example, on 101-point attitude scales (sometimes called feeling thermometers), what exactly do 76, 77, and 78 mean? Even for 11- or 13-point scales, people may be hard pressed to define the meaning of the scale points" (p. 270). The simulations reported on in Figure 3 represent the idea that our discriminating powers are more limited than that implied by the 101-point scale. It assumes that true scores vary across the discrete set of 11 integers on a 0 to 10 scale, rescaled as 0, 10, 20, …, 80, 90, and 100. Everything else in the set up is the same as in the earlier simulations, but in this case the 101-point scale is incapable of improving on the 11-point scale in terms of the information it conveys. Any observed score that deviates from 0, 10, 20, …, 80, 90, or 100 is only adding noise to the data.

Not surprisingly, the results presented in Figure 2 and 3 show reliability peaking with the 11-point scales and declining with the 101-point scale.[14] The extra noise induced by the 101-point scale, very evident in the bottom panel of Figure 2, outweighs the ability of the 101-point scale to capture more true variability in scores, leading it to suffer in a reliability comparison with coarser scales, consistently.

[14] To simplify, we just present one set of true score and random error distributions in these simulations, as described in the figure titles.

In Figure 3, the fact that the 101-point scale differs only from the 11-point scale in its ability to pick up noise means it is consistently more troubled by random error variance than is the 11-point scale (bottom panel) and is never more reliable than the 11-point alternative (top panel). In contrast to what we see in Figure 2, however, here the reliability disadvantage of the longer scale diminishes as the randomness in the responses increases. In all other respects the results from simulations depicted in Figure 2 and 3 support the conclusions drawn from those described in Figures 1a-1d.

The value of these simulations can be found in their ability to discipline our thinking about how and why reliability may vary with the coarseness of a scale in a single-item measure of an attitude. Whether there are any lessons to be drawn for practice is less clear. Even if one wished to choose a scale length solely on the basis of its reliability—which would be a mistake—it remains an open question as to which scale length that would be. Figures 2 and 3 showed that the 101-point scale could be made to lose to the 11-point scale in a reliability contest under fairly plausible assumptions, even though the assumptions underlying the contrary results in Figures 1a-1d were not blatantly far-fetched. Later we turn to an empirical investigation of the question, which may yield a clearer verdict on the matter.

Simulation Results: Validity and Effect Estimation

Although the validity of a measure is not affected by random errors in measurement, it is affected by systematic errors in measurement. The rounding induced by using a coarsened measure—or, equivalently, by using a collapsed version of an otherwise continuous measure—will invariably introduce systematic error and, hence, lower the validity of the measure. That systematic error, in turn, will lead to biases in the estimation of causal effects and other quantities of interest. Krieg (1999, p. 764) summarized the gloomy outlook thusly: "Scale coarseness can affect the mean, variance, covariance, correlation coefficient, and the reliability of scores. Both the numerator and denominator of the correlation coefficient can be affected. The biases can vary as a function of the number of scale points and the number of items in a scale, as well as the mean and variance of the quantities to be measured. Different rules for assigning values to scale points can produce different biases." Our simulations provide a means

to illustrate the systematic error introduced by rounding and its consequences for estimation, and pave the way for a parallel demonstration based upon an analysis of the experimental data.

The systematic errors found in coarsened scales are affected by the exact values assigned to the scale points. For these simulations we scored all scales to range from 0 (most conservative/Republican) to 100 (most liberal/Democratic), which is comparable to 0-1 coding conventionally used by scholars working with attitude measures. Thus, for example, the 2-point scale took on the values of 0 and 100 while the 5-point scale took on the values of 0, 25, 50, 75, and 100. Recall that rounding error is the difference between the observed score and the score that would have been obtained without rounding—the "actual score", which is the sum of the true score plus random error. Figure 4 shows how the average rounding error (on the Y-axis) covaries with true scores (on the X-axis) across scales of varying length. This set of results builds in the assumption that true scores follow a uniform distribution over the 0-100 interval.[15]

To understand the patterns that one sees in Figure 4, consider the example of a 2-point scale. True scores less than 50 are rounded to code 0 while true scores greater than 50 are rounded to code 100. This generates underestimates of increasing magnitude as true scores range from 0 to 50, and then overestimates of decreasing magnitude as true scores range from 50 to 100. The rare score landing exactly at the cutpoint (50) will end up scored either at 0 or at 100, about half the time each. The pattern of underestimation followed by overestimation repeats as the scale length (and number of cutpoints) grows. The result is that in each case rounding error is positively correlated with true scores, but to a diminishing extent as the scale length increases. For the simulations underlying the plots in Figure 4, the correlation between the true scores and the rounding errors are .50, .34, .25, .19, .16, .13, .09, and -.00, for the 2, 3, 4, 5, 6, 7, 11, and 101-point scales, respectively.

[15] The distribution of random errors does not materially affect these results, but this particular set of simulations assumed normal random error, SD=5. The decision about whether or not to censor the data (round observed scores below 0 to 0, and scores above 100 to 100) does matter to the details, but not to the general point of these or subsequent results. Here, we did not censor scores. If we do censor scores then the average rounding error at 0 and 100 becomes a bit more extreme than that shown in Figure 4.

The fact that rounding in Y produces errors that increase with Y means that estimates of causal effects will tend to be biased away from zero.[16] To illustrate this, we constructed a simulation in which true scores were a function of an experimental X (treatment=1 vs. control=0). We specified an effect size of 30, centering the distributions of true scores on 50 (with the treatment group centered at 65 and the control group centered at 35). As usual, we added random error to the true scores (normally or uniformly distributed, with a variable SD) to yield actual scores, and then rounded the actual scores to scales of 2, 3, 4, 5, 6, 7, 11, and 101 points. We then regressed Y on X for each of our 1000 samples (n=1000 each). The average of the unstandardized regression coefficients are shown in Figure 5. As the figure shows, the coefficients obtained when analyzing coarsened scales will tend to overestimate the causal effect, with the extent of the bias diminishing as the scale lengthens and the quantity of rounding error diminishes.[17] In these results the bias is minimal when rounding to an 11-point scale.

These simulations build in the assumption that true scores vary continuously, but what if true scores only vary discretely, perhaps along an 11-point scale, as discussed above? Two points are important. First, if true scores fall along an 11-point scale, then using a coarser scale (e.g., 5 or 7 points) will have the negative consequences we have already discussed. Second, if one makes the mistake of using a more fine-grained scale, like the 101-point scale, then that will certainly have the effect of introducing more noise into your measure. But it should not result in more systematic measurement error; certainly, none of the rounding error problems we have discussed are introduced. As such, erring in the direction of too fine a measure is the lesser sin.[18]

---

[16] Alternative coding schemes produce systematic rounding error that can bias estimates in other ways. For instance, coding items to the midpoint of the range for each scale segment, rather than fully to the ends of the 0-100 scale, will produce rounding error that biases estimates of causal effects towards zero. In this case, rounding errors are negatively correlated with true scores.

[17] Bias in the coefficients depends on where the true scores of treatment and control subjects fall relative to the cutpoint that determines how scores are rounded. That is why the estimated effect sizes fluctuate rather than shift monotonically as scale length increases. For the 5-point scale, when random error is minimal, most of the control and treatment subjects get assigned to option 2 (scored 25) and option 4 (scored 75), respectively, which yields an estimated coefficient close to 50. By contrast, for the 4 point scale most of the control subjects choose option 2 (score 33.3) while most of the treatment subjects choose option 3 (score 66.7), which yields an estimated coefficient closer to 33.3. This kind of dynamic affects the simulated results and, presumably, actual ones as well.

[18] Given space limitations, we cannot provide a detailed assessment of how scale length affects inferences when attitudes serve as an independent variable. However, the reliability and validity problems with coarsened scales that

Experimental Study Design

We fielded a two-wave survey on Amazon's Mechanical Turk in October, 2013. A total of 1676 respondents completed the first wave, while 1346 completed both the first and second wave, resulting in a successful recontact rate of 80.3%. Respondents were compensated $0.25 for completing the first survey and $0.50 for completing the second. When to field the second wave of a test-retest survey is always an important design question. The goal is to let enough time pass so that respondents will not remember their wave 1 responses and yet not so much time that true change is expected to have occurred between waves. One to three weeks between waves has been recommended in the literature. We recontacted all respondents approximately 8 days after they completed the first survey. Reminder emails were sent every 48 hours if the respondents had not yet completed the follow-up. Overall, 95% of wave 2 responses were provided by 14 days after the initial survey.

The sample of respondents looks similar to most Mechanical Turk (MTurk) samples (Berinsky et al. 2012, Buhrmester et al. 2011, Mason and Suri 2011). Table 1 describes its composition on select political and demographic variables. Overall, the sample was young (average age=32), 58% male, and 75% white. Democrats and Liberals outnumbered their counterparts on the right by a ratio of about 2 to 1. Only about 30% reported having cast a vote in the 2012 presidential election, which no doubt reflects the youthful quality of the sample as well as the respondents' mild level of interest in politics (only 26% said they follow politics most of the time).

Working with this MTurk sample means that our test-retest correlations are probably lower than they would be with a representative sample of the U.S., because the attitudes of young people tend to be less crystallized (Sears 1986). Also relevant to the generalizability of the study's findings is the fact that the field period coincided with the federal government shutdown (October 1-October 16, 2013) and its aftermath. This turmoil may have encouraged more true attitude change than we would see in a more

---

we have identified will also thwart valid inference in that context. One difference between the two is that both random and systematic error in attitude measures will bias causal inferences when they serve as independent variables, while only systematic error will bias causal inferences when they serve as dependent variables.

placid period.[19] Such changes, which will be regarded as errors in a test-retest analysis, are more likely to be felt in more fine-grained measures, as our simulations have demonstrated. Thus, the differences in test-retest reliabilities by scale length may be attenuated relative to what one would find in a different context.

Survey respondents were randomly assigned in wave 1 of the survey to receive one of twelve different attitudinal scale formats. These scales were repeated in the second wave of the experiment so that test-retest reliability could be calculated for all respondents on identical scale formats. Table 2 lists the various scale formats included in the study as well as the number of respondents in each condition. The Appendix provides the exact visual layout of each of the scale formats. The scales varied in length (2, 3, 4, 5, 6, 7, 11, and 101 points), the number and wording of labels, and the response format (radial vs. slider vs. feeling thermometer).

Our focus in the analysis to come is on a subset of the conditions that allow for the cleanest comparisons, those that contrast the 11-point and 101-point sliders to the 2, 3, 4, 5, 6, and 7-point scales. The 2-7 point scales had 2-7 labels, respectively, while the 11 and 101-point sliders provided verbal labels for the endpoints and middle-position that match those used for the 3, 5, and 7 point scales (very unfavorable, very favorable, neither favorable nor unfavorable), while also providing eleven numerical labels—0, 1, 2,…8, 9, 10 (11-point) or 0, 10, 20, …80, 90, 100 (101-point). We also consider the consequences of including or excluding a middle option, focusing here on the 3-point vs. 2-point scale, the 5-point vs. the 4-point scale, and the 7-point vs. the 6-point scale.[20]

[19] Across the 16 political targets included in the survey (detailed below), there were 5 statistically significant differences in means across waves ($p < .05$). These targets—John Boehner, Paul Ryan, the Tea Party, Republicans, and Conservatives—all were rated more negatively in wave 2. However, the largest substantive difference, for John Boehner, was only -0.035 (with scales coded 0-1), only 3.5% of the scale's length.

[20] Thus, 4 of the 12 experimental conditions/scale formats do not figure into the comparisons we discuss. We exclude the 11-point radial scale because while it can be compared to the 11-point slider and to the 2 through 7-point radial scales, we have no radial version of the 101-point scale. We exclude the 101-point slider with 7 labels because the only clean comparison is with the 7-point scale with 7 labels. We exclude the 3 and 7-label feeling thermometer scales—which showed respondents the thermometer graphic used by the American National Election Studies and asked them to supply a number indicating their response—because these conditions can only be cleanly compared to the 101-point sliders.

The survey instrument in both waves was identical, except for the fact that we included more demographic questions in the wave 1 survey. Respondents were asked to provide their attitudes towards 24 different targets, presented in three batteries of 8 each. One battery asked about 8 celebrities,[21] chosen to vary in their notoriety as well as in their presumed partisan affiliation. Eight of the targets were individual political figures,[22] four Democrat, four Republican, and varied in prominence. The remaining eight of these targets were political groups,[23] with four typically aligned with the Republican party and four typically aligned with Democrats.

To keep any order effects constant across waves, respondents always completed the celebrity battery first, then the political figure battery, and then the political groups battery. The order of targets within each battery was randomized and then held constant across both waves. The survey also include measures of celebrity knowledge and political knowledge, interest in each of these domains, party identification, ideological self-identification, and 2012 vote choice.

The strategy of gauging attitudes toward a large and diverse set of political and non-political figures means that the data can either be analyzed item-by-item or can be stacked and analyzed as a whole. To clarify this difference, assume for the moment that we had just 1000 2-wave responses. With the data arranged to provide wave 1 and wave 2 variables (columns) for each of our 1000 cases (rows), we could provide 24 test-retest correlations for each of the 8 scale length/experimental conditions of interest (2, 3, 4, 5, 6, 7, 11, and 101), and average some or all of the results. Alternatively, we could stack the data so that we have two variables, wave1 and wave 2 attitude responses (columns), and 24,000 observations (rows). With this dataset structure we would simply calculate the wave-1-wave-2 correlations for each of the 8 experimental groups.[24] While stacking overstates the number of

[21] Lindsay Lohan, Ben Affleck, Meryl Streep, Rosie O'Donnell, Jay-Z, Tina Fey, Justin Bieber, and Clint Eastwood.
[22] John Boehner, Bill Clinton, George W. Bush, Joe Biden, Nancy Pelosi, Barack Obama, Paul Ryan, and Mitt Romney.
[23] Democrats, Gays & Lesbians, Labor Unions, Tea Party, Big Business, Republicans, Conservatives, and Liberals.
[24] Item fixed effects need to be included. Without item fixed effects, test-retest correlations will be inflated to the extent that there are stable mean differences across items. Even if T1 and T2 responses were perfectly uncorrelated for each item, we would observe a positive test-retest correlation within the stacked data.

independent observations we have, which matters for statistical significance testing, that can be accounted for at the analysis stage. More importantly, stacking has two virtues. First, it diminishes the skew in the frequency distributions, which is valuable since skew attenuates test-retest (Pearson) correlations, especially with coarse measures (Wylie 1976). Second, with stacked data one can take into account certain kinds of response sets with models that incorporate individual-level fixed effects. We analyze the data both ways in what follows.

Experimental Results: Descriptive Statistics and Response Times

Figures 6a and 6b provide descriptive statistics on the 24 attitude measures in the survey, broken down by scale length (2, 3, 4, 5, 6, 7, 11, or 101). All scales were recoded to the 0-1 interval, with 1 indicating high favorability. The targets are rank ordered based on the mean (Figure 6a) and standard deviation (Figure 6b) of the ratings, averaged across the two survey waves.

Respondents varied quite dramatically in their favorability towards the various celebrity and political targets. Not surprising in light of the Democratic leanings of our sample, Democratic targets are rated more favorably than Republican targets. However, the top- and bottom-rated targets across most scale lengths are celebrities (Tina Fey and Justin Bieber, respectively). As the scale lengthens, the means gravitate towards the center across all targets. However, they do not do so monotonically—as middle options are added and taken away, the means shift slightly up and down, reflecting the ability for respondents to record their attitude at the midpoint. More polarizing and well-known figures such as Barack Obama record the highest standard deviations, while a number of celebrities found at the bottom of figure 6b record the lowest, as they are nearly universally liked or disliked by the MTurk respondents. The standard deviations of political targets are highest in the 2 and 3 point scales, as fewer options results in a more polarized distribution, given our 0 to 1 scaling.

Respondents were timed as they recorded their favorability ratings of the targets. Figure 7 records the average amount of time respondents spent answering the questions in the two political attitude batteries. The data show a strong, positive relationship between scale length and the time respondents spent determining and recording their attitudes. Of note is the relatively little variation in response time

by type of target (politicians, groups). Averaging across all 16 items and both waves, the time spent ranged from 17.3 seconds (2.2 seconds per question) for the 2-point scale to 29.5 seconds (3.7 seconds per question) for the 101-point scale. The biggest jumps in time taken occurred when moving from the 2 to the 3-point scale (2.7 seconds for the battery), from the 7 to the 11-point scale (2.3 seconds), and from the 11 to the 101-point scale (3.2 seconds). While this hearty relationship seems to suggest that respondents are taking their task seriously, spending more time when asked to provide a more discerning assessment, it cannot tell us whether the extra time taken with longer scales levels the measurement error playing field across lengths.

Figure 8 shows response-time results for the high and low political knowledge halves of the sample.[25] Notably, and as one would expect, high knowledge respondents tended to take more time reporting their attitudes across all scale lengths. However, the biggest gaps between the high and low knowledge groups are found for the 11 and 101-point scales. The differences average out to 4.4 seconds for those two scales, compared to 2.0 seconds for the six coarser scales. Low information respondents could be less comfortable making fine-grained judgments than high information respondents, and thus quickly gravitate to simple benchmarks, like the scale mid-point (which they do populate much more frequently), or simply be more frustrated with the task's difficulty and thus make less effort (Krosnick and Presser 2010).[26]

Experimental Results: Test-Retest Reliability Coefficients

Of primary interest to us is how the test-retest reliabilities vary by the length of the scale. Figure 9 presents these results for the full sample, both when analyzing the stacked data and when simply averaging the 24 test-retest coefficients calculated from the un-stacked data (labeled "Averaging" in Figure 9). A first point to make is that the stacked correlations are slightly higher than the average un-

[25] The political knowledge scale consisted of five factual questions about politics, displayed in the Appendix. The average number of correct answers was 3.2. High knowledge respondents are those who answered three or more correctly, while low knowledge respondents are those who answered two or fewer correctly.
[26] Funke, Reips, and Thomas (2011) did not vary scale length, but found that less educated respondents were especially likely to have trouble with sliders compared to scales with radial buttons.

stacked correlations, as expected, but the pattern of variation across scale length is the same in the two instances. A second point is that the differences by scale length are small in magnitude, with reliabilities ranging from a low of about .74 (for the 2-point scale) to a high of about .84 (for the 11-point scale). As we have already argued, however, the observed test-retest results will overstate the true reliability of the most coarsened scales, so the actual difference in reliability across scale length is greater than these results signify.

The most important result in Figure 9 is the relatively poor performance of the 101-point scale—a slider with three labels and 11 numerical labels (0, 10, 20, …, 80, 90, 100)—compared to the 11-point scale—also a slider, with the same system of labeling. Figure 10, which presents the stacked reliability results for the 16 political items, separately for high and low knowledge subsamples, shows the same pattern. Estimated reliability is lower for the low-knowledge group across the board, save for when they are responding to the 2-point scale, but the attitude measures are more reliable for both groups when they are answering on an 11-point scale than when they are answering on the 101-point scale.

The other main pattern in Figures 9 and 10 is that the test-retest reliabilities grow steadily, though not quite monotonically, as one moves from 2 to 11 scale points. Although the observed differences are not great, as already noted, there is no evidence in these results that one would be better off, reliability-wise, by using an attitude scale less than 11 points long.

Figure 11 rearranges some of the data to highlight the comparison between the even-length scales (2, 4, 6) and their counterparts that offer a middle option (3, 5, 7). Although most of the scholarly consternation over the question of whether to use middle options involves Likert scales (see, e.g., Johns 2005, O'Muircheartaigh, Krosnick and Helic 1999), researchers using rating scales must decide whether to offer one. Unfortunately, our data do not allow us to extend the comparison to 10 vs. 11-point scales, but the comparison of 2 vs. 3, 4 vs. 5, and 6 vs. 7 all point in the same direction: excluding a midpoint carries the price of a small loss in reliability.

Because the study design included multiple attitudinal targets, we can extend the analysis to take into account response sets, i.e., the tendency of respondents to use a particular point or range of the scale.

Some individuals may repeatedly gravitate toward the mid-point, if one is available. Others may tend to avoid negative evaluations, while still others may embrace them. We analyzed response sets in two ways. In the first, we simply added respondent fixed effects to the models yielding the test-retest reliabilities based on the stacked data. Doing so essentially factors out the mean rating provided by each respondent across the 24 different targets prior to calculating the test-retest reliability. Figure 12 shows the results by scale length, with the baseline (no respondent fixed effects) also shown for comparison.[27] Not surprisingly, the reliability coefficients tend to be (slightly) lower when respondent fixed effects are included. Still, the pattern seen earlier, with the 101-point scale performing worse than the 11-point scale, remains.

Our second strategy was to try and identify respondents who appeared to be engaging in what Krosnick (1991, 1999) calls non-differentiation: failing to give differentiated responses to the different items in a battery. Non-differentiation is thought to arise when respondents are giving little thought or attention to their responses ("satisficing"). We constructed three alternative measures to classify non-differentiators, varying the strictness of the classification. The first identified a respondent as a non-differentiator if he or she gave the exact same response when evaluating the 8 targets in at least one of the three batteries (celebrities, politicians, groups). The second classified a respondent as a non-differentiator if he or she gave the exact same response when evaluating the two targets in one or more of the following pairs: Obama and Romney, Clinton and Bush, Democrats and Republicans, Liberals and Conservatives (at either wave). The third classified a respondent as a non-differentiator if his or her response had the same valence for the two targets in one or more of those pairs (e.g., both Obama and Romney rated positively).

Figure 13 reports the rate of non-differentiation by scale length using these measures. Using the strictest measure, non-differentiation rates are very low and tend to decline with scale length. Respondents were less likely to give all targets in a battery the same rating when they had 101 points to

[27] Another way to specify the model would allow a unique respondent response set—and fixed effect—for each battery (celebrities, politicians, political groups). Doing so yields results that are similar to those given in Figure 12.

choose from than when they had 2 or 3 points to choose from. This is what one would expect based solely on the number of choices presented to respondents. More interesting are the results obtained when using the other two, looser measures of non-differentiation. In both, non-differentiation rates tend to decline with scale length up until the 11-point scale is reached, and then to increase with the 101-point scale. Using the strict rule applied to 4 target pairs, non-differentiation rates climb from 32% for the 11-point scale to 39% for the 101-point scale. Using the directional rule applied to 4 target pairs, the rate of non-differentiation increases from 57% (11-point) to 78% (101-point). This is further evidence that the 101-point scale is overly demanding, at least for some respondents.

The higher rate of non-differentiation found for the 101-point scale relative to the 11-point scale could, in principle, be responsible for its weaker reliability performance. But, as Figure 14 shows, even when we analyze scale reliability separately among differentiators and non-differentiators, the poorer performance of the 101-point scale persists. This figure shows test-retest reliabilities on all 24 items (from the stacked data, with item fixed effects) for the two sets of respondents using our strict and directional 4-target measures. It shows, first, that reliabilities are systematically lower among non-differentiators compared to differentiators. The lack of attention and effort in responding that led to their classification as non-differentiators is equally evident in the greater noise in their responses. At the same time, the 101-point scale remains less reliable than the 11-point scale for three of the four groups, especially so among the non-differentiators.

In understanding these reliability results it is useful to refer back to the simulation findings. When it comes to the 101-point vs. 11-point scale comparison, it is clear that the greater random error in responses to the former outweighs the greater variation in true responses it is able to provide. This could occur because the 101-point instrument itself induces more measurement error into the responses—because, for example, respondents find it too difficult and time-consuming to complete carefully—as illustrated by the Figure 2 simulation. It could also arise because all or some respondents are incapable of making the fine-grained distinctions the 101-point scale enables, as illustrated by the Figure 3 simulations. It is not that respondents failed to make use of the many response options that the 101-point

scale afforded them, however. Figure 15 depicts the (stacked) distribution of responses across the 16 political items for the 11 and 101-point scales, with higher numbers indicating more liberal/Democratic responses. The MTurk respondents readily abandoned the labeled 0, 10, 20, …, 80, 90, 100 options when given the opportunity to express a more fine-grained response. But, giving them that opportunity yielded more noise than it did new information.

Experimental Results: Estimation Biases

The study design did not include an experimental X expected to affect our attitudinal responses, so we cannot perform an analysis of bias in estimating causal effects that exactly parallels our simulations of the same. However, we can show how the relationship between an observational X and our attitudinal measures (Ys) varies by scale length. Given what the simulations have shown, we would expect the association to diminish as the scale lengthens. Our observational analyses produce exactly this result. The estimated "effect" of an observational X diminishes with scale length.

Figure 16 shows these results for two observational Xs: a dummy variable indicating whether the respondent identified as Democrat (X=1) or not (X=0), and a dummy variable indicating whether the respondent identified as liberal (X=1) or not (X=0). The regression analysis made use of stacked data on the 16 political targets, coded to range from 0 (most conservative/Republican response) to 1 (most liberal/Democratic response). We ran regressions with each dummy variable in turn, including item fixed effects. As the figure shows, the magnitude of the regression coefficients drops by about by about one-third as the scale moves from being 2-points (coefficients on the order of .35) to 101-points (coefficients on the order of .24) in length.

Notably, the regression coefficients for the 11-point and 101-point scale are very similar. What is not similar between that pair of scales is the R-squared obtained from each regression, shown in Figure 17. Although the fit of the regression tends to improve as the scale lengthens from 2 to 11-points (the recurring anomaly of the 6-point scale set aside), it then worsens when moving to the 101-point scale. This pair of results regarding the 11-point vs. 101-point scale—with comparable "effect" estimates but worse fit for the 101-point scale—jives perfectly with our previous conclusions. There is more noise in

the 101-point scale than in the 11-point scale, hence the worse fit obtained when trying to explain it by any X. Yet the two scales are comparable in terms of their systematic error, yielding similar findings when estimating how they are affected by that X.[28]

By contrast, the results for the coarser scales contain both larger regression coefficients and lower R-squared values than those found for the 11 point scale. Since the 0 to 1 coding of the attitude measures yields rounding errors that are positively correlated with the true scores, as our simulations showed, it follows that the regression coefficients for the coarser scales are biased upward. And since the coarser attitude scales are also less reliable, it follows that there is more noise in the regression equation seeking to explain them. Once again, the evidence provides good reason to avoid using coarsened scales.

Conclusion

The simulations we designed and reported on in this paper provide a useful framework for understanding the scale length problem, particularly in their partitioning of random error and rounding error in survey responses. They make clear how the well-known, if oft-ignored, problems introduced when one collapses a continuous measures coincide with the problems introduced when one opts for a coarse measure in the first place. The results reveal both the value and the limits of fine-grained measures. Although fine-grained measures are ordinarily superior, in reliability and validity terms, their value diminishes if measurement error grows with scale length or if respondents are only able to differentiate between a smaller set of points than the scale presents. In such cases there is nothing to be gained by moving to a longer/finer measure.

Because of the sensitivity of these conclusions to the underlying assumptions of the simulations, we conducted a parallel empirical study that helps clarify the practical effect of choosing different scales varying in length. Attitudes were seemingly quite reliable across scale lengths, with even the 2-point scale achieving a test-retest reliability of .74. However, as the simulations demonstrated, test-retest correlations overstate the true reliability, particularly for short (2-4 point) scales, which must be taken into

[28] The fact that the 101-point scale is less reliable than the 11-point scale is more problematic when attitudes serve as X. As is well-known, random error in X tends to bias effect estimates, attenuating bivariate associations.

account when making comparisons. Our experimental results further showed that while reliability steadily increased as the number of scale points grew from 2 to 11, it then decreased among both high and low knowledge respondents when moving to a 101-point scale. Non-differentiating, seemingly careless, responses were also more likely to appear with the 101-point than with the 11-point scale. While the 101-point scale offers a more fine-grained measure, the random error it introduced more than exceeded the gain in true variation (if any) that it provided.

We are not ready to advise researchers to abandon the 101-point attitude scale in favor of the 11-point scale. The costs of employing an unduly coarse measure—in terms of lowered reliability and validity, with the associated biases and power limitations in statistical estimation—are too significant. The poor showing of the 101-point scale in our study may be anomalous. And as our simulation results have suggested, fine-grained measures are more valuable when random error in measurement is limited, likely to be the case when people care a lot about the topic or target of the attitude. The fact that our MTurk subjects were unusually low in political engagement may have contributed to the poor showing of the 101-point scale. Likewise, it is too soon to conclude that the 11-point attitude scale should always be chosen over the 7-point scale. Too fine a measure carries its own costs. We need independent evidence confirming these findings based on more diverse samples and taking into account multiple survey modes. In the meantime, scholars should err in the direction of seeking more fine-grained rather than less fine-grained measures. Although measures that are needlessly coarse and those that are needlessly fine-grained each have their problems, the latter is not vulnerable to the rounding error that can seriously confound inferences.

References

Achen, C. H. (1975). Mass political attitudes and the survey response. *The American Political Science Review*, pages 1218–1231.

Aguinis, H., Pierce, C. A., and Culpepper, S. A. (2009). Scale coarseness as a methodological artifact correcting correlation coefficients attenuated from using coarse scales. *Organizational Research Methods*, 12(4):623–652.

Alwin, D. F. (1992). Information transmission in the survey interview: Number of response categories and the reliability of attitude measurement. *Sociological Methodology*, pages 83–118.

Alwin, D. F. (1997). Feeling thermometers versus 7-point scales which are better? *Sociological Methods & Research*, 25(3):318–340.

Alwin, D. F. and Krosnick, J. A. (1991). The reliability of survey attitude measurement the influence of question and respondent attributes. *Sociological Methods & Research*, 20(1):139–181.

Andrews, F. M. (1984). Construct validity and error components of survey measures: A structural modeling approach. *Public opinion quarterly*, 48(2):409–442.

Berinsky, A. J., Huber, G. A., and Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon. com's mechanical turk. *Political Analysis*, 20(3):351–368.

Blok, H. and Saris, W. E. (1983). Using longitudinal data to estimate reliability. *Applied psychological measurement*, 7(3):295–301.

Bollen, K. A. and Barb, K. H. (1981). Pearson's r and coarsely categorized measures. *American Sociological Review*, pages 232–239.

Brudvig, S. (2007). *From Coarse to Fine and Weak to Strong: The Impact of Scale Granularity and Rating Strength on the Ability of K-means to Recover True Cluster Structure*. Doctoral Dissertation.

Buhrmester, M., Kwang, T., and Gosling, S. D. (2011). Amazon's mechanical turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6(1):3–5.

Christian, L. M., Parsons, N. L., and Dillman, D. A. (2009). Designing scalar questions for web surveys. *Sociological Methods & Research*, 37(3):393–425.

Cicchetti, D. V., Shoinralter, D., and Tyrer, P. J. (1985). The effect of number of rating scale categories on levels of interrater reliability: A monte carlo investigation. *Applied Psychological Measurement*, 9(1):31–36.

Cook, C., Heath, F., Thompson, R. L., and Thompson, B. (2001). Score reliability in webor internet-based surveys: Unnumbered graphic rating scales versus likert-type scales. *Educational and Psychological Measurement*, 61(4):697–706.

Couper, M. P., Tourangeau, R., Conrad, F. G., and Singer, E. (2006). Evaluating the effectiveness of visual analog scales a web experiment. *Social Science Computer Review*, 24(2):227–245.

Cox III, E. P. (1980). The optimal number of response alternatives for a scale: A review. *Journal of marketing research*, pages 407–422.

Elkins, Z. (2000). Gradations of democracy? empirical tests of alternative conceptualizations. *American Journal of Political Science*, 44:193–200.

Ferrando, P. J. (2003). A kernel density analysis of continuous typical-response scales. *Educational and psychological measurement*, 63(5):809–824.

Finstad, K. (2010). Response interpolation and scale sensitivity: Evidence against 5-point scales. *Journal of Usability Studies*, 5(3):104–110.

Funke, F., Reips, U.-D., and Thomas, R. K. (2011). Sliders for the smart: Type of rating scale on the web

interacts with educational level. *Social Science Computer Review*, 29(2):221–231.

Green, D. P. and Palmquist, B. (1994). How stable is party identification? *Political behavior*, 16(4):437–466.

Hofmans, J., Theuns, P., and Mairesse, O. (2007). Impact of the number of response categories on linearity and sensitivity of self-anchoring scales. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 3(4):160–169.

Hulbert, J. (1975). Information processing capacity and attitude measurement. *Journal of Marketing Research*, pages 104–106.

Johns, R. (2005). One size doesn't fit all: Selecting response scales for attitude items. *Journal of Elections, Public Opinion & Parties*, 15(2):237–264.

Krebs, D. (2012). The impact of response format on attitude measurement. In *Methods, Theories, and Empirical Applications in the Social Sciences*, pages 105–113. Springer.

Krieg, E. F. (1999). Biases induced by coarse measurement scales. *Educational and Psychological Measurement*, 59(5):749–766.

Krosnick J. A. Response strategies for coping with the cognitive demands of attitude measures in surveys. Applied Cognitive Psychology 5 (1991) 213–236.

Krosnick J. A. Survey research. Annual Review of Psychology 50 (1999) 537–567.

Krosnick, J. A. and Berent, M. K. (1993). Comparisons of party identification and policy preferences: The impact of survey question format. *American Journal of Political Science*, pages 941–964.

Krosnick, J. A., Boninger, D. S., Chuang, Y. C., Berent, M. K., and Carnot, C. G. (1993). Attitude strength: One construct or many related constructs? *Journal of personality and social psychology*, 65(6):1132.

Krosnick, J. A. and Fabrigar, L. R. (1997). Designing rating scales for effective measurement in surveys. *Survey measurement and process quality*, pages 141–164.

Krosnick, J. A. and Presser, S. (2010). Question and questionnaire design. *Handbook of survey research*, 2:263–314.

Lehmann, D. R. and Hulbert, J. (1972). Are three-point scales always good enough? *Journal of Marketing Research*, pages 444–446.

Lozano, L. M., Garcìa-Cueto, E., and Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 4(2):73–79.

Mason, W. and Suri, S. (2012). Conducting behavioral research on amazon's mechanical turk. *Behavior research methods*, 44(1):1–23.

Maydeu-Olivares, A., Kramp, U., Garcìa-Forero, C., Gallardo-Pujol, D., and Coffman, D. (2009). The effect of varying the number of response alternatives in rating scales: Experimental evidence from intra-individual effects. *Behavior research methods*, 41(2):295–308.

McDowell, I. (2006). *Measuring health: a guide to rating scales and questionnaires*. Oxford University Press.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological review*, 63(2):81.

O'Muircheartaigh, C. A., Krosnick, J. A., and Helic, A. (2001). *Middle alternatives, acquiescence, and the quality of questionnaire data*. Irving B. Harris Graduate School of Public Policy Studies, University of Chicago.

Owen, S. V. and Froman, R. D. (2005). Why carve up your continuous data? *Research in nursing & health*, 28(6):496–503.

Pasek, J. and Krosnick, J. A. (2010). Optimizing survey questionnaire design in political science: insights from psychology. *Oxford handbook of American elections and political behavior*, pages 27–50.

Pearse, N. (2011). Deciding on the scale granularity of response categories of likert type scales: The case of a 21-point scale. *Electronic Journal of Business Research Methods*, 9(2).

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., and Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *Journal of applied psychology*, 88(5):879.

Preston, C. C. and Colman, A. M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta psychologica*, 104(1):1–15.

Ramsay, J. O. (1973). The effect of number of categories in rating scales on precision of estimation of scale values. *Psychometrika*, 38(4):513–532.

Schaeffer, N. C. and Presser, S. (2003). The science of asking questions. *Annual Review of Sociology*, pages 65–88.

Schwarz, N., Knäuper, B., Hippler, H.-J., Noelle-Neumann, E., and Clark, L. (1991). Rating scales numeric values may change the meaning of scale labels. *Public Opinion Quarterly*, 55(4):570–582.

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of personality and social psychology*, 51(3):515.

Sherif, M. and Hovland, C. I. (1961). Social judgment: Assimilation and contrast effects in communication and attitude change.

Shively, W. P. (1998). *The craft of political research*. Prentice Hall.

Srinivasan, V. and Basu, A. K. (1989). The metric quality of ordered categorical data. *Marketing Science*, 8(3):205–230.

Visser, P. S. and Krosnick, J. A. (1998). Development of attitude strength over the life cycle: surge and decline. *Journal of personality and social psychology*, 75(6):1389.

Weng, L.-J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement*, 64(6):956–972.

Wylie, P. B. (1976). Effects of coarse grouping and skewed marginal distributions on the pearson product moment correlation coefficient. *Educational and Psychological Measurement*, 36(1):1–7.

Zaller, J. (1992). *The nature and origins of mass opinion*. Cambridge University Press.

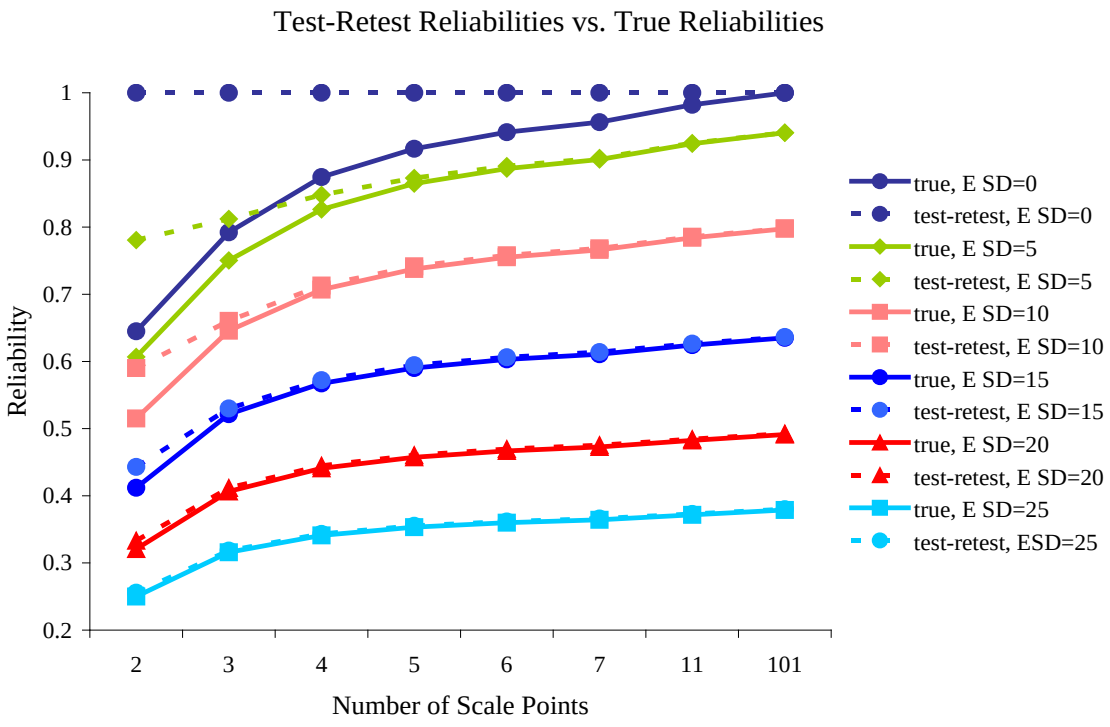
Table 1. Selected Characteristics of Experimental Sample

Party ID (with leaners in party)	Democrat 59.6%	Independent 18.0%	Republican 21.3%		
Ideology (collapsed to 3 point)	Liberal 57.3%	Moderate 22.9%	Conservative 18.9%		
Voted in 2012	Yes 28.8%	No 69.2%			
Highest Educational Degree	H. School 33.9%	Associate's 15.6%	Bachelor's 38.4%	Master's 7.2%	Doctorate 2.9%

Table 2. Experimental Conditions and Sample Size

Experimental Condition	Labels on Scale	Wave 1 Completions	Wave 2 Completions	Completion Rate
2-point radial	2	147	119	81.0%
3-point radial	3	138	109	79.0%
4-point radial	4	145	127	87.6%
5-point radial	5	130	101	77.7%
6-point radial	6	123	103	83.7%
7-point radial	7	154	129	83.8%
11-point radial	3	145	116	80.0%
11-point slider scale	3	130	107	82.3%
101-point slider scale	3	133	109	82.0%
101-point slider scale	7	137	109	79.6%
101-point feeling thermometer	3	144	113	78.5%
101-point feeling thermometer	7	136	105	77.2%

Figure 1a: Simulated Results, Normal True, Normal Error



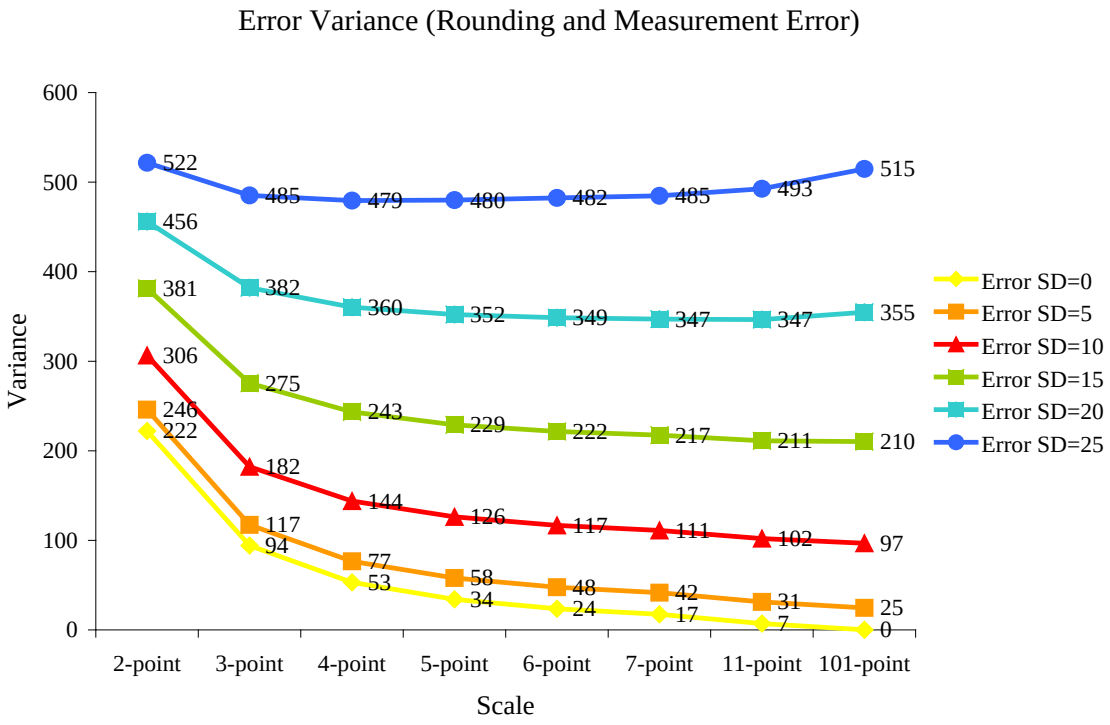


Figure 1b: Simulated Results, Normal True, Uniform Error

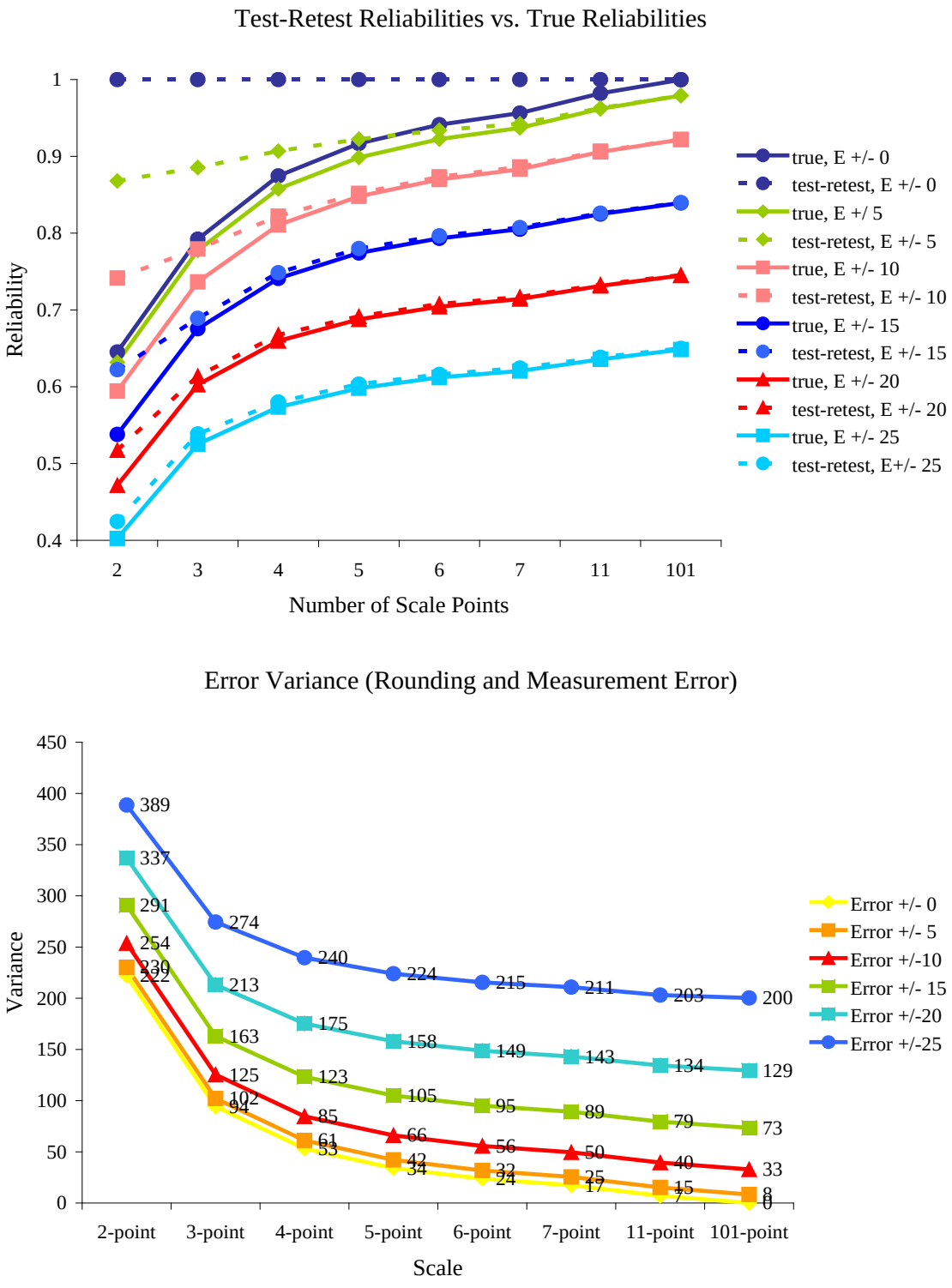
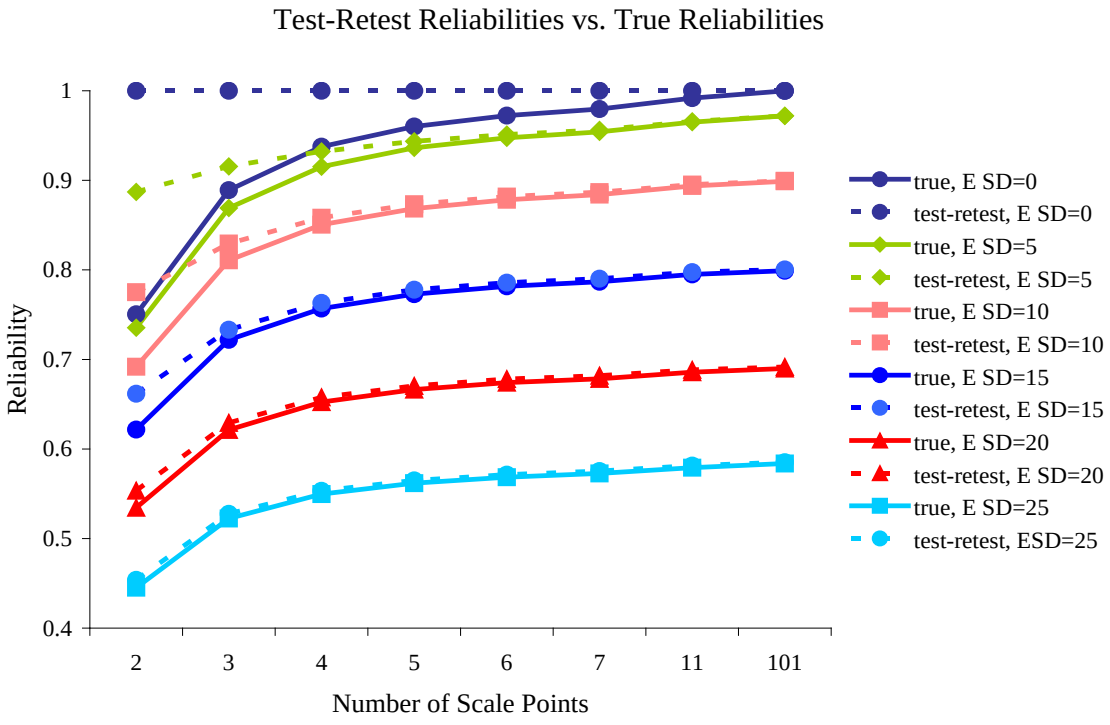


Figure 1c: Simulated Results, Uniform True, Normal Error



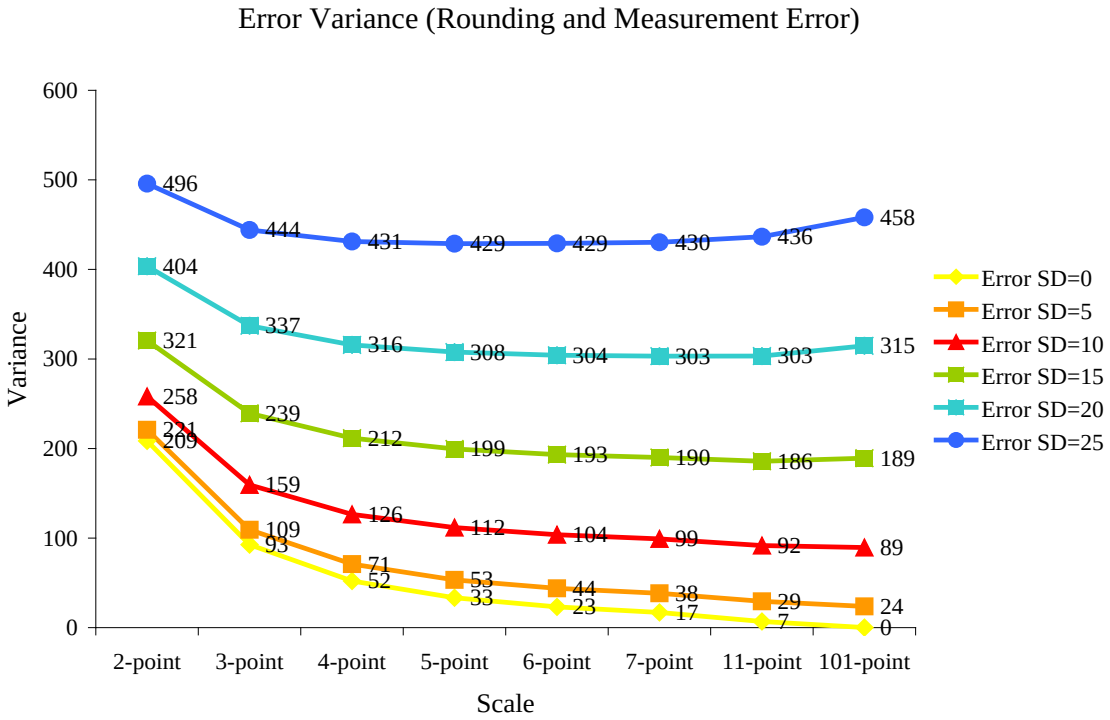
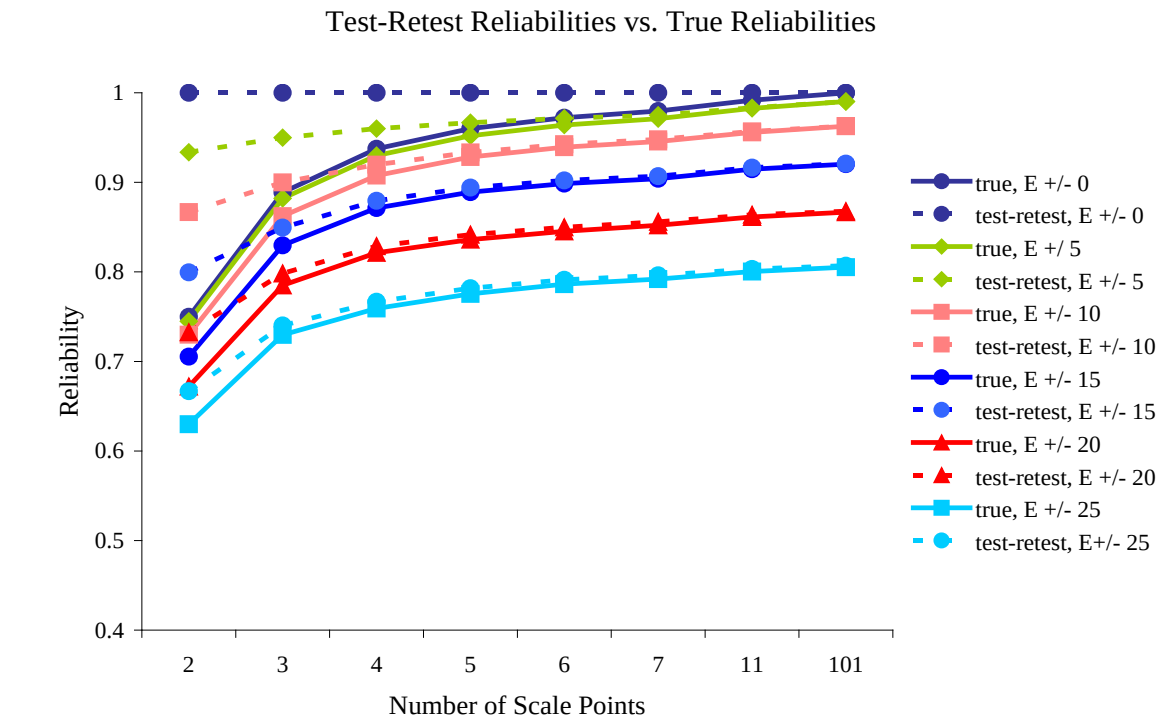


Figure 1d Simulated Results, Uniform True, Uniform Error



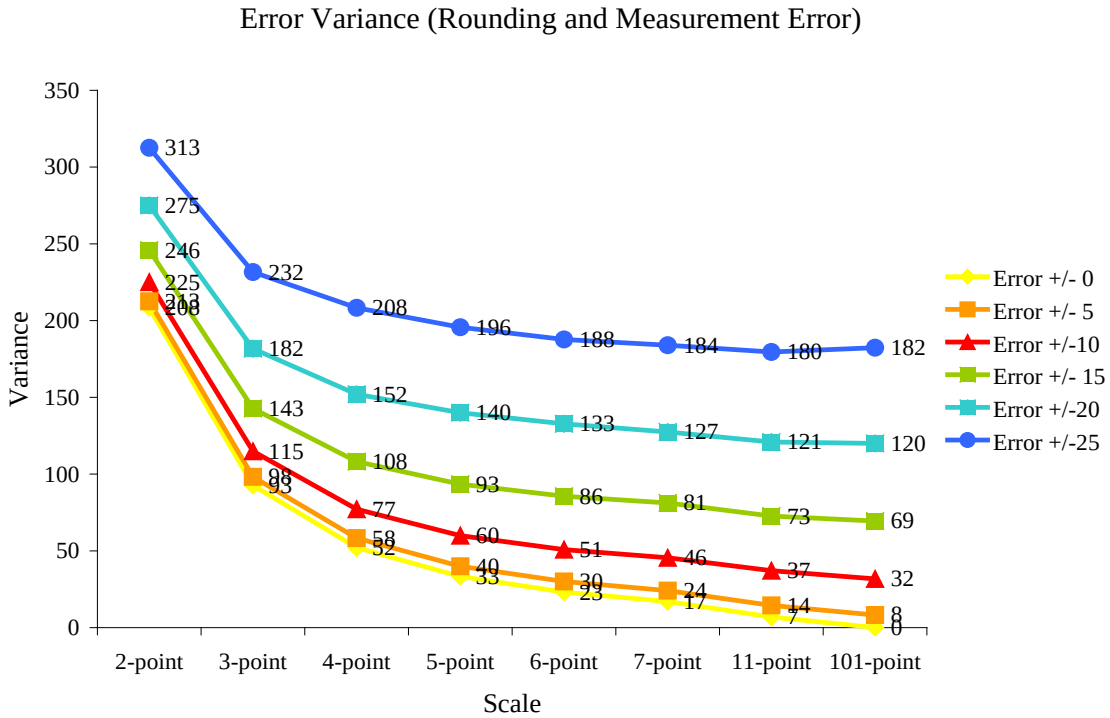
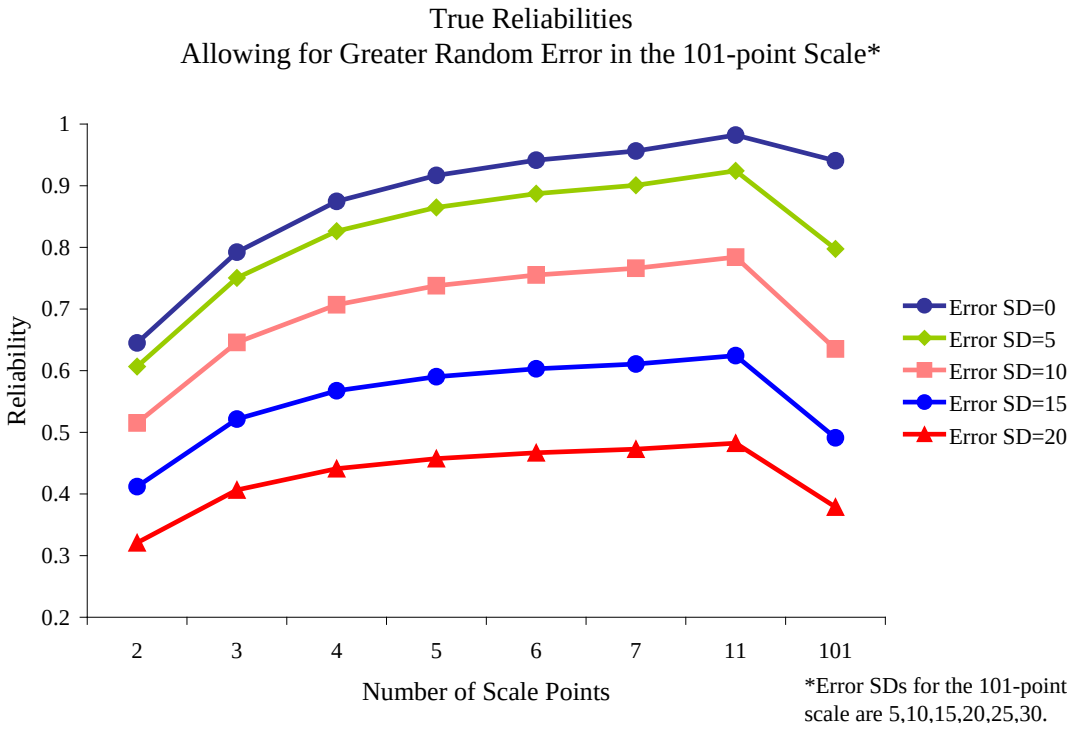


Figure 2: Simulated Results (True=N, Error=N) with more Error in the 101-point Scale



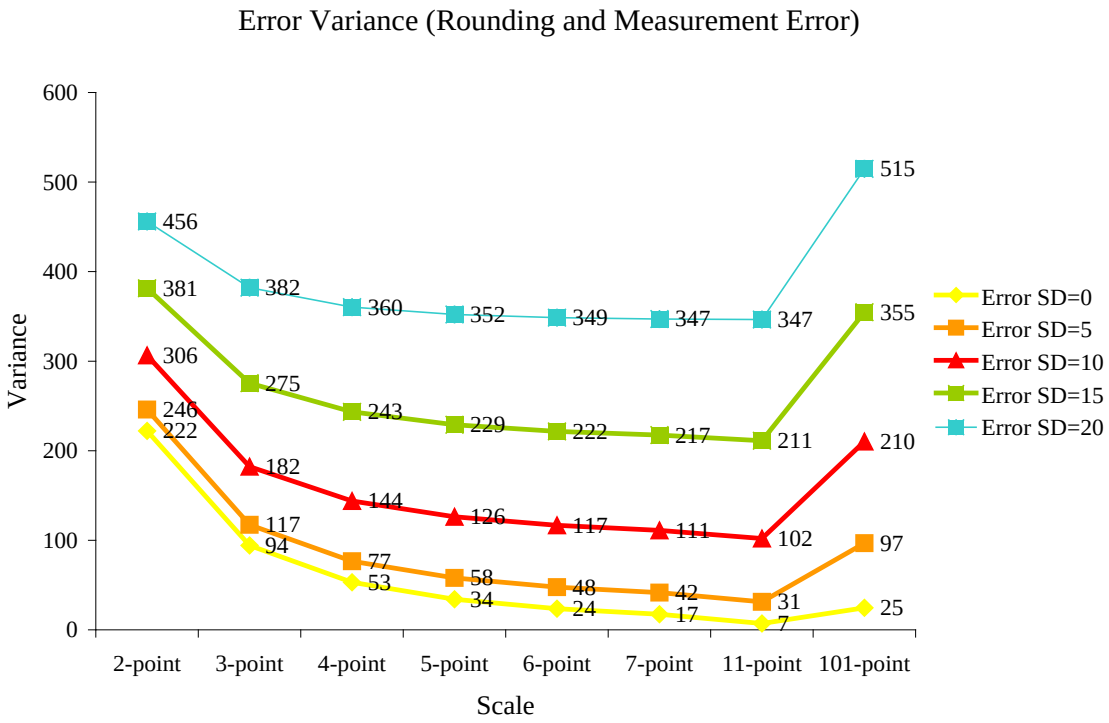
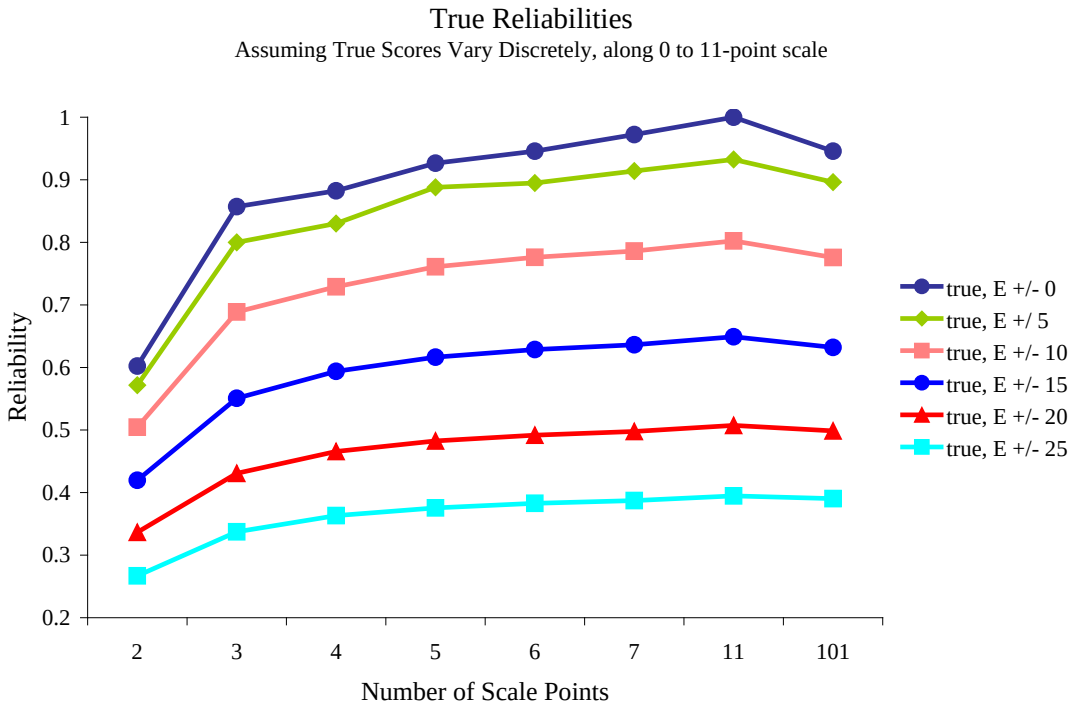
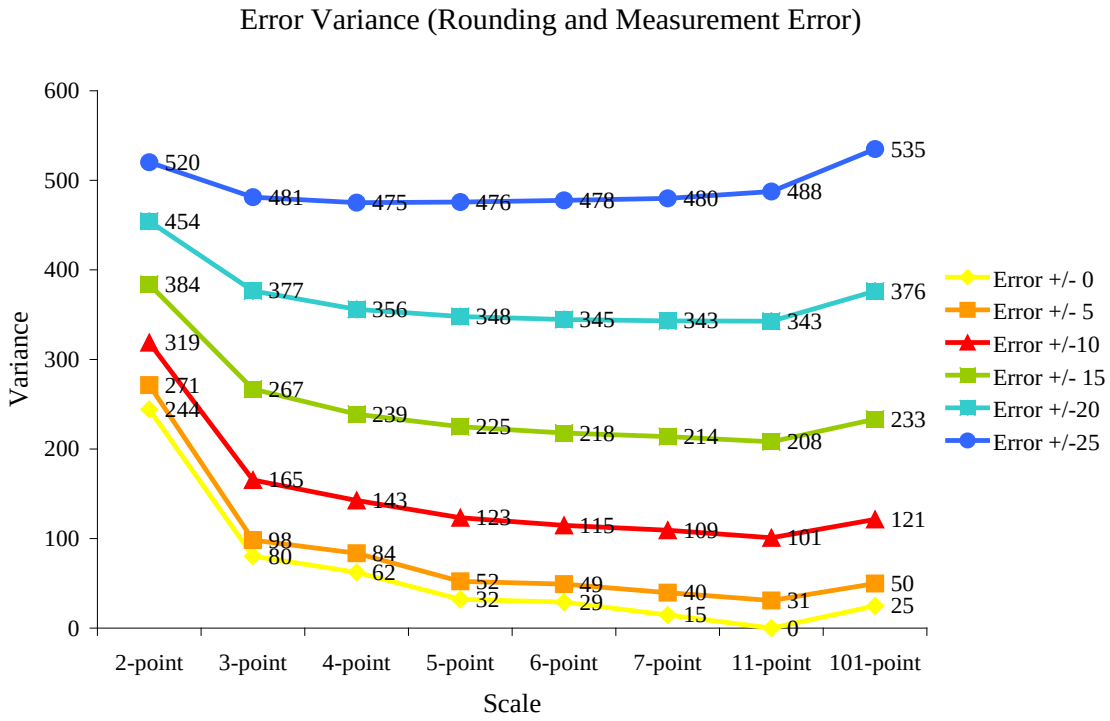


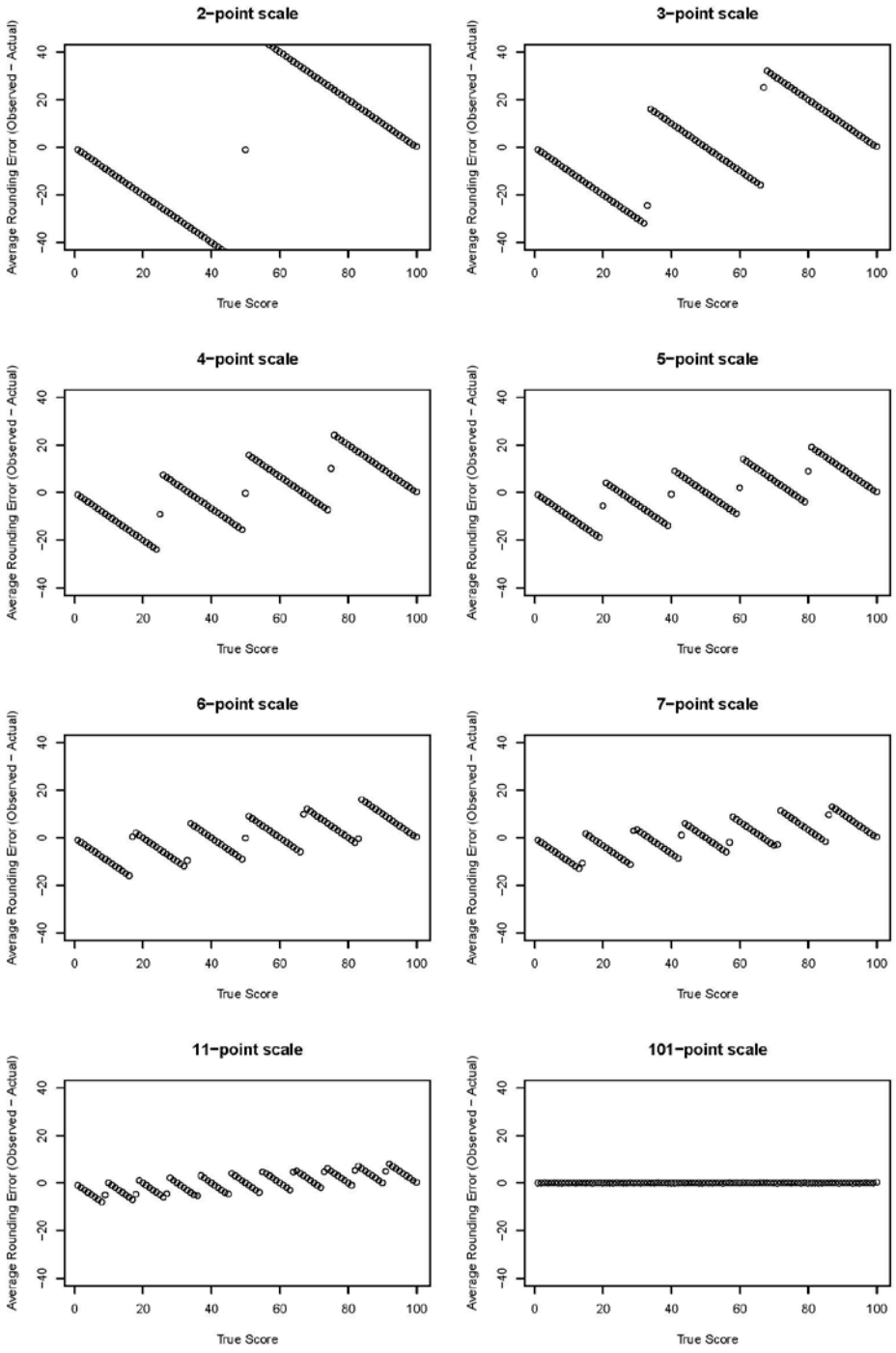
Figure 3: Simulated Results, True = Peaked,[29] Discrete, 11-point Scale, Error=Uniform





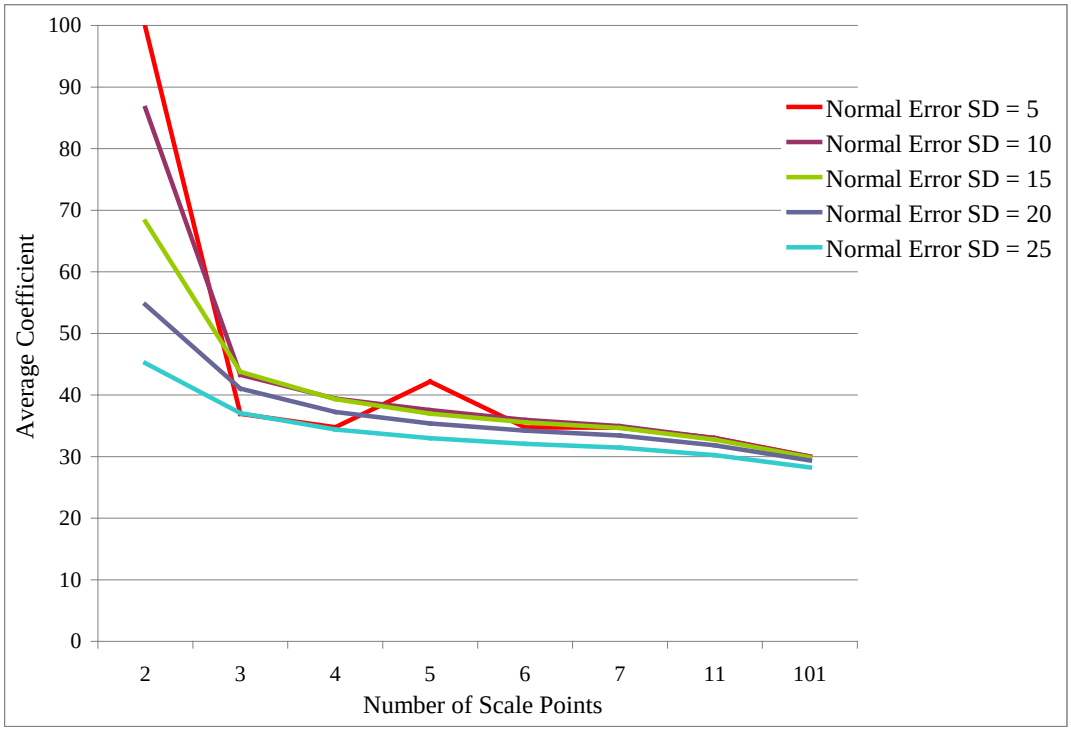
[29] Probabilities = {.025,.05,.075,.1,.125,.25,.125,.1,.075,.05,.025}

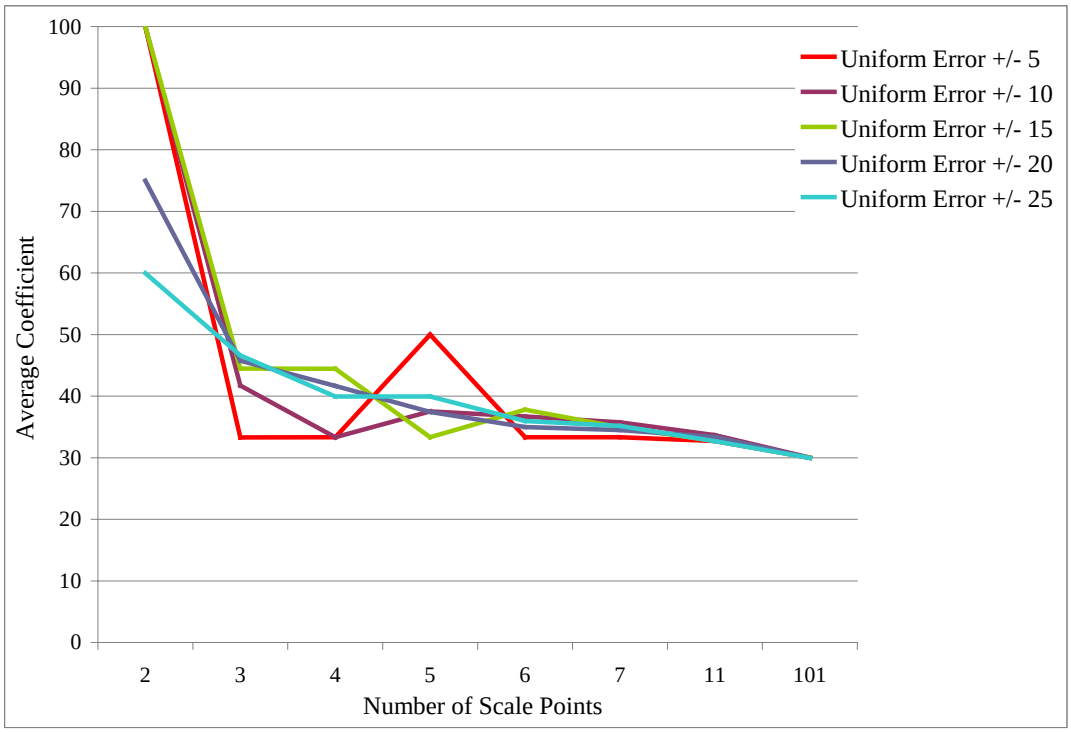
Figure 4: How Rounding Error Varies by True Scores



Note: The Y axis is the average rounding error across simulations, where rounding error is the difference between the observed score and the "actual score" (true score plus rounding error). Observed scores were coded to range from 0 (minimum) to 100 (maximum). For example, the two point scale was coded 0,100, while the three point scale was coded 0, 50, 100.

Figure 5: Bias in Causal Effect Estimation Introduced by Rounding Error





Note: Entries are unstandardized regression coefficients representing the estimated effect of an experimental dummy variable on the attitude measure. True effect=30. See text for further details.

Figure 6a: Mean Ratings by Target and Scale Length

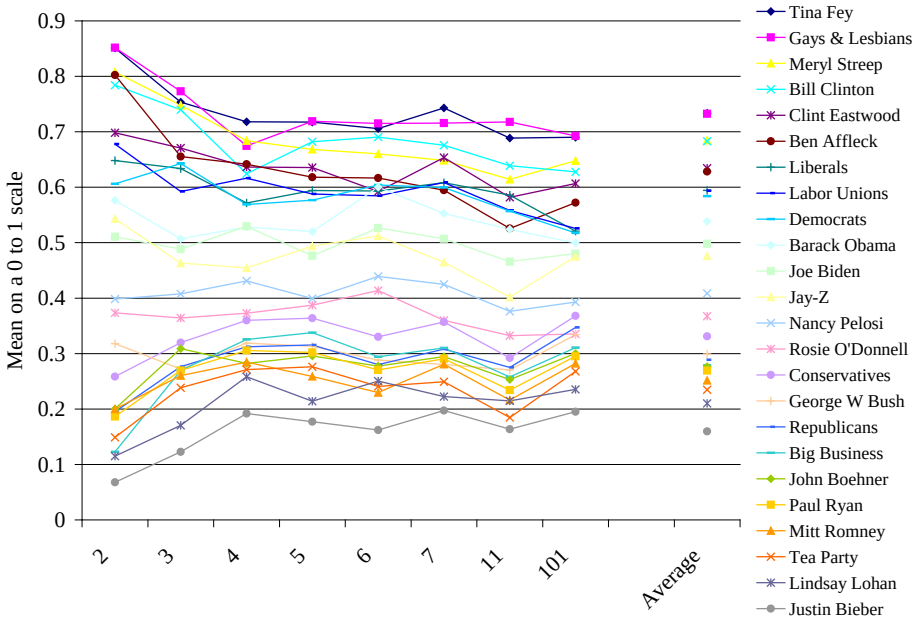


Figure 6b: Standard Deviations of Ratings by Target and Scale Length

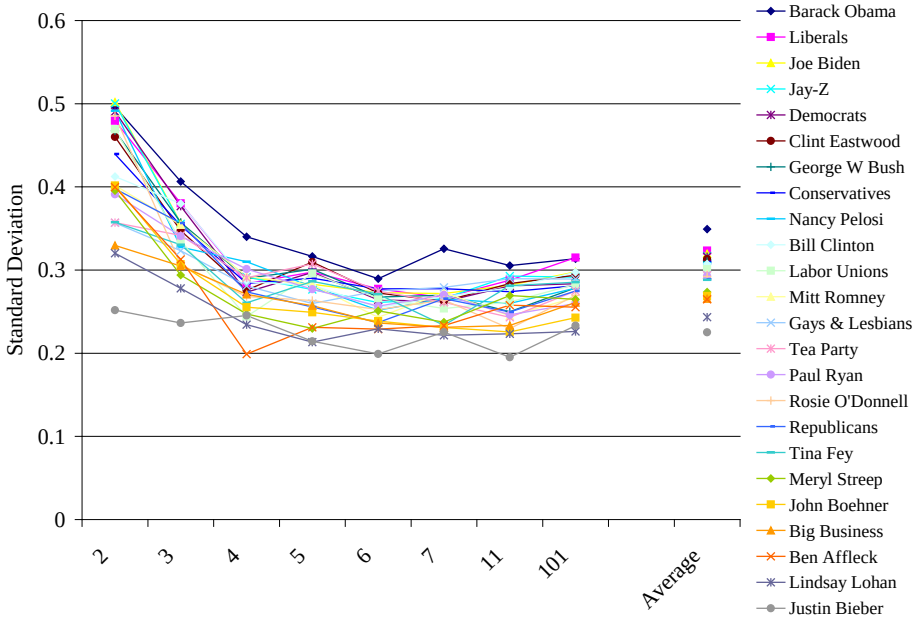


Figure 7: Response Times by Battery and Scale Length

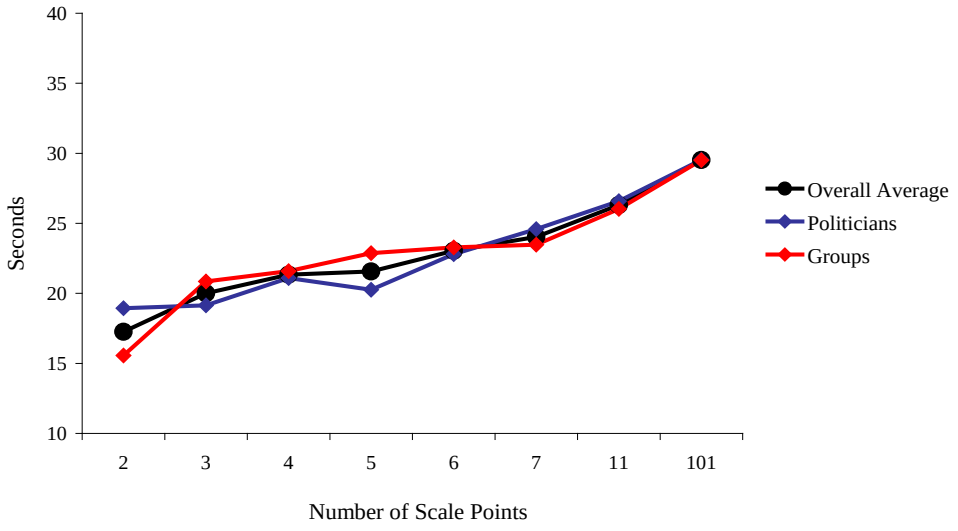
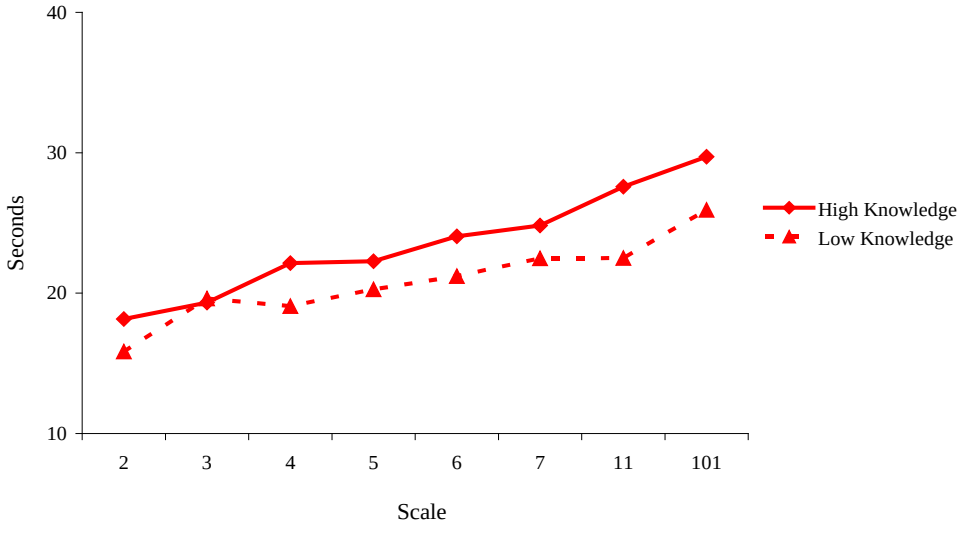
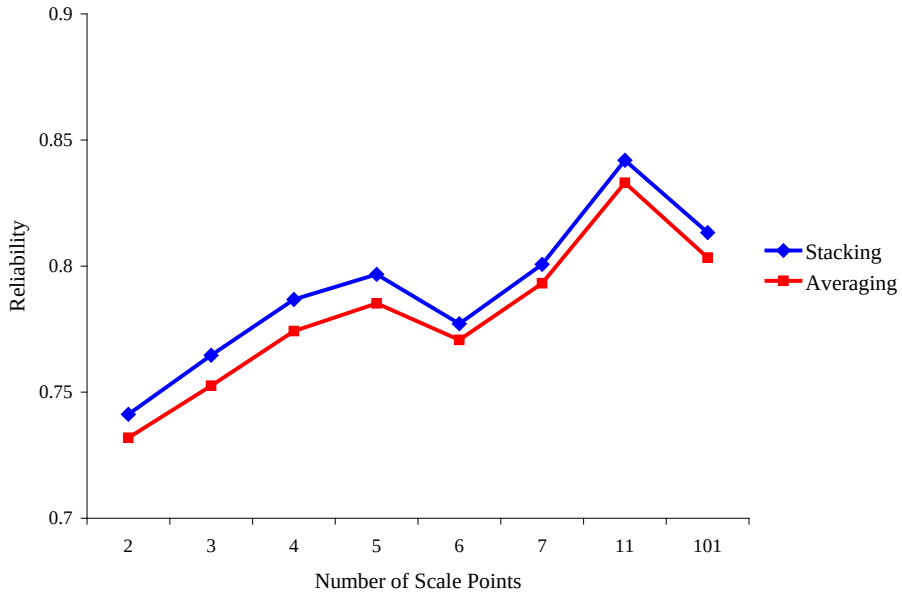


Figure 8: Response Times by Political Knowledge and Scale Length



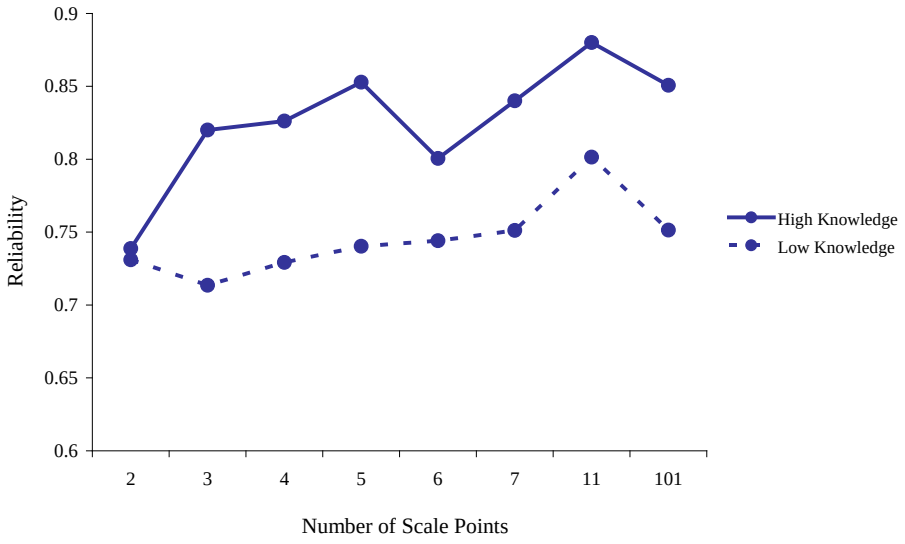
Note: Quantities shown average across the 8 politician and 8 group items.

Figure 9: Test-Retest Reliabilities by Scale Length



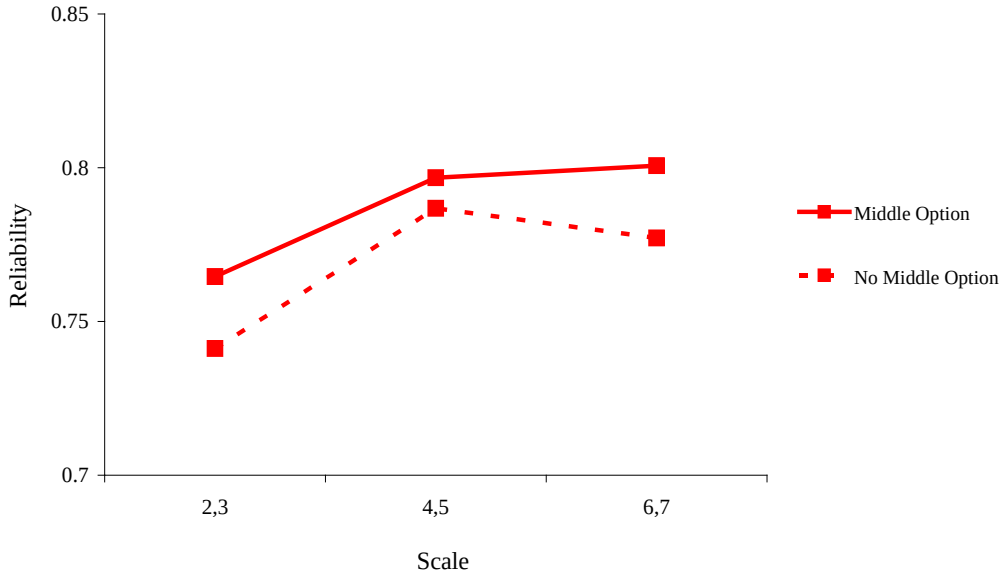
Note: Shown are test-retest correlations averaged across all 24 items ("Averaging") or as given by an analysis of the stacked data ("Stacking"). Analyses of the stacked data included item fixed effects.

Figure 10: Test-Retest Reliabilities by Political Knowledge and Scale Length



Note: The figure shows test-retest correlations obtained from analyses of the stacked data on attitudes toward the 16 political targets; analyses included item fixed effects. Results based on averaging the item-specific reliability coefficients are comparable. See text for details on the classification of respondents into low vs. high political knowledge.

Figure 11: Test-Retest Reliabilities by the Availability of a Middle Alternative



Note: Results from analysis of stacked data are shown. These are drawn from Figure 9.

Figure 12: Test-Retest Reliabilities With and Without Respondent Fixed Effects

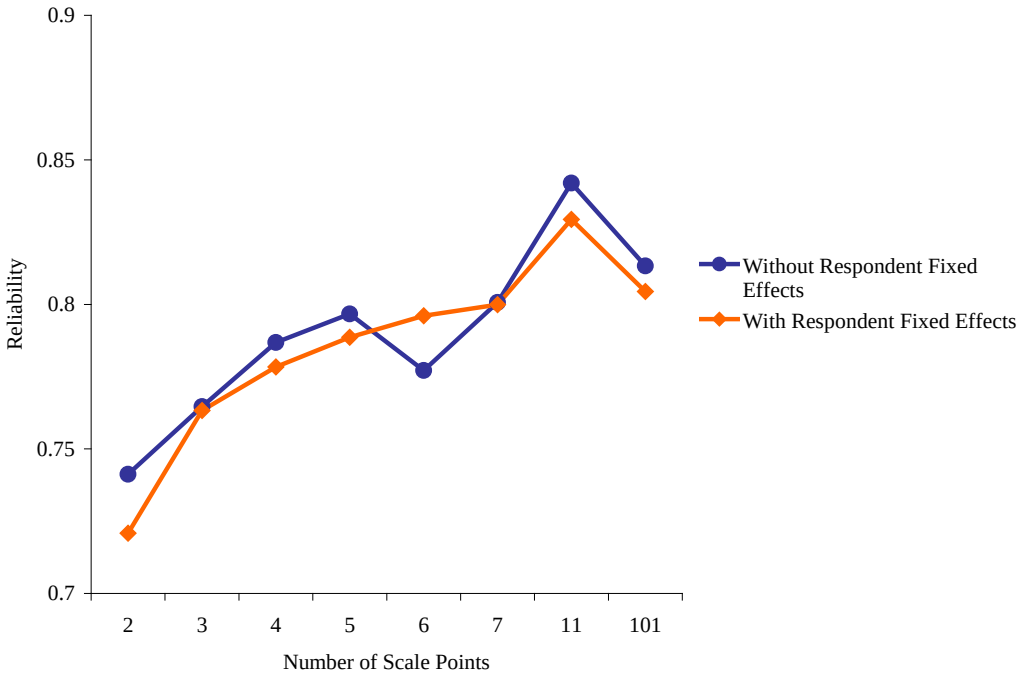
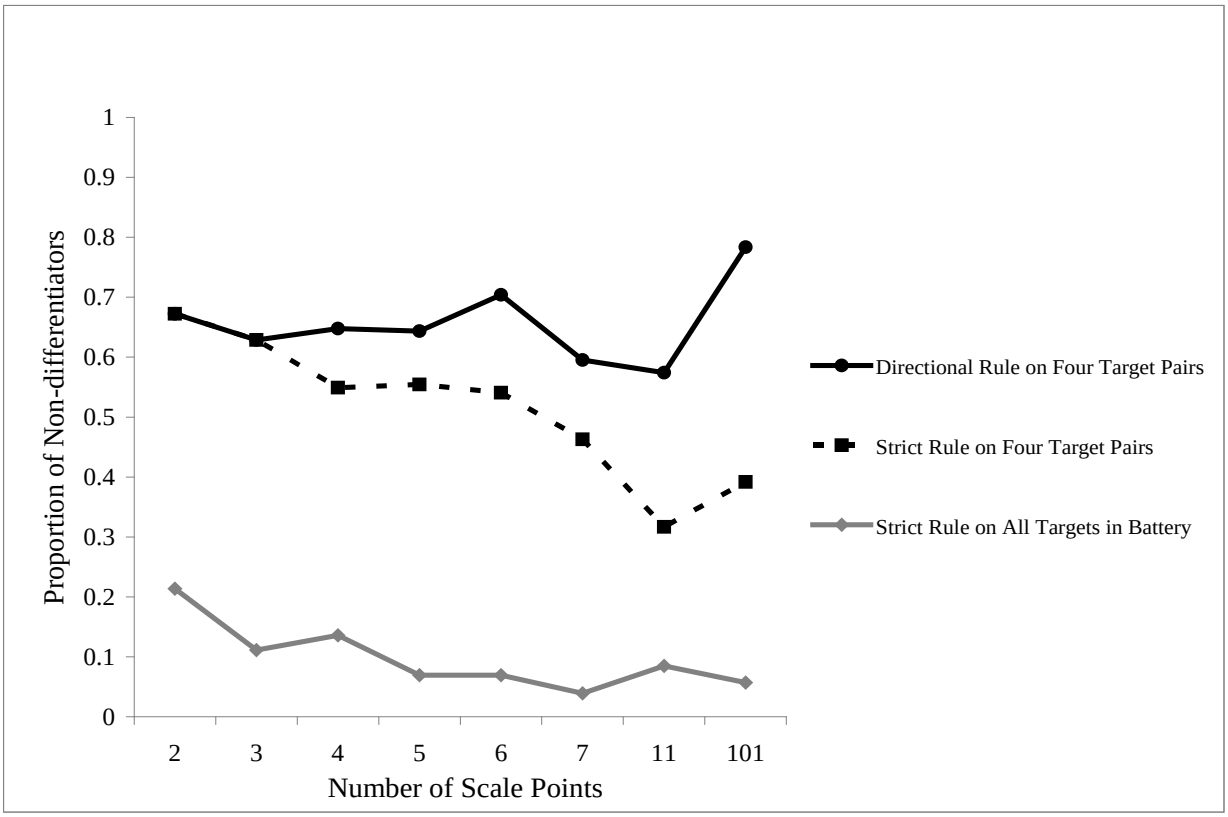
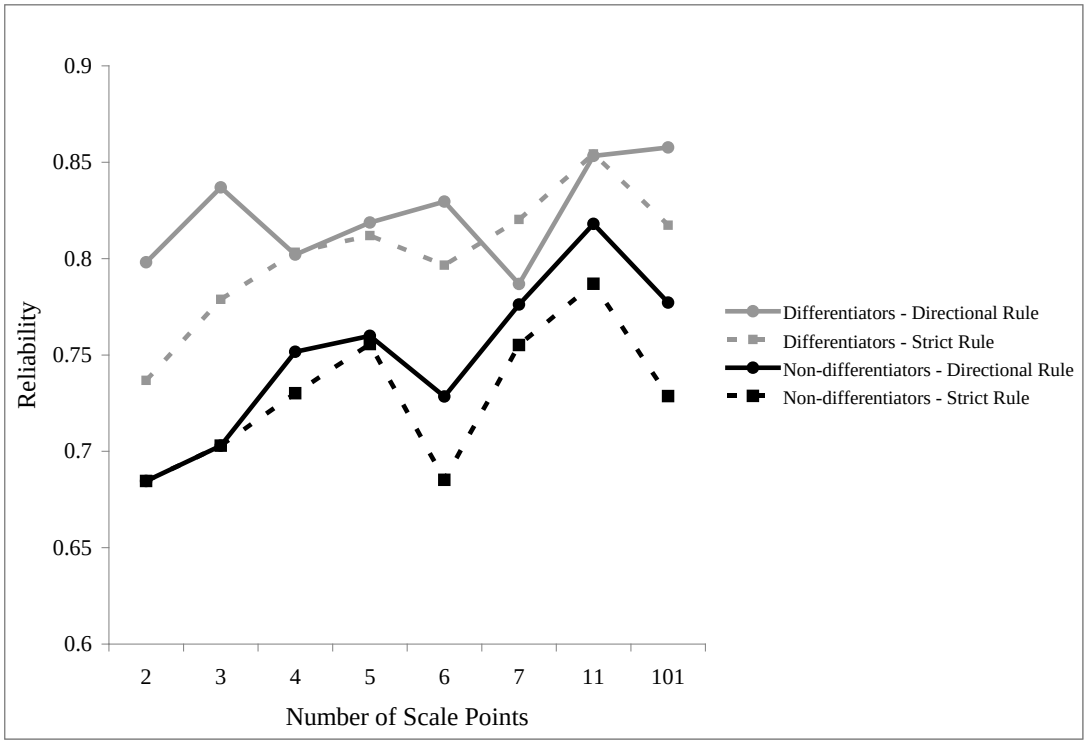


Figure 13: Non-differentiation Rates by Scale Length



Note: Following the "strict rule on four target pairs," non-differentiators are those who gave the same score to the two individuals/groups in one or more of these pairs: Obama and Romney, Clinton and Bush, Democrats and Republicans, Liberals and Conservatives. Following the "directional rule on four target pairs," non-differentiators are those who gave scores with the same direction (positive or negative) to the two individuals/groups in one or more of the pairs. Following the "strict rule on all targets in battery," non-differentiators are those who gave the same score to all eight targets in at least one of the three batteries (celebrities, politicians, political groups).

Figure 14: Reliability among Differentiators and Non-Differentiators, by Scale Length



Note: All analyses are based on the stacked data and include item fixed effects.

Figure 15: Distribution of Ratings on 11-point and 101-point Scales

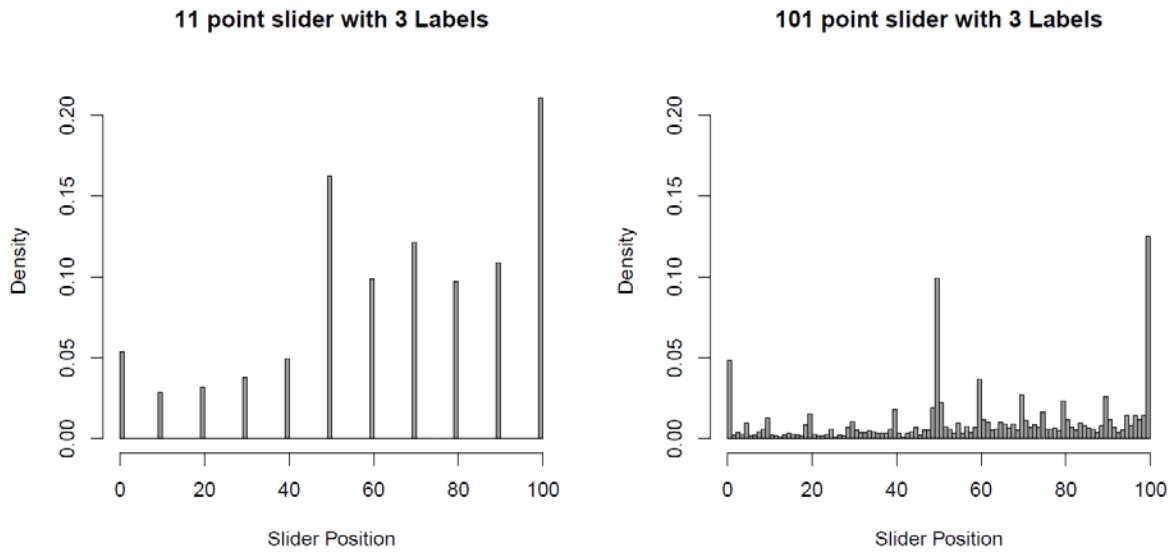
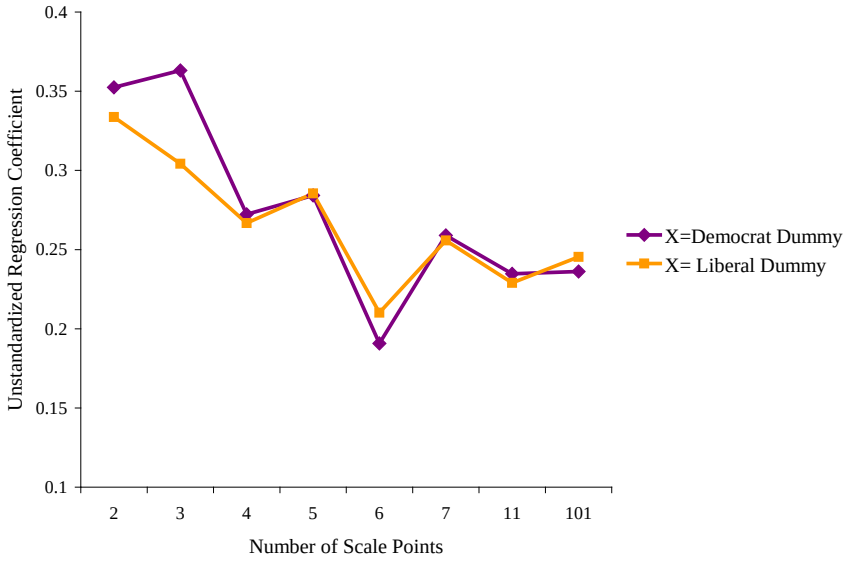
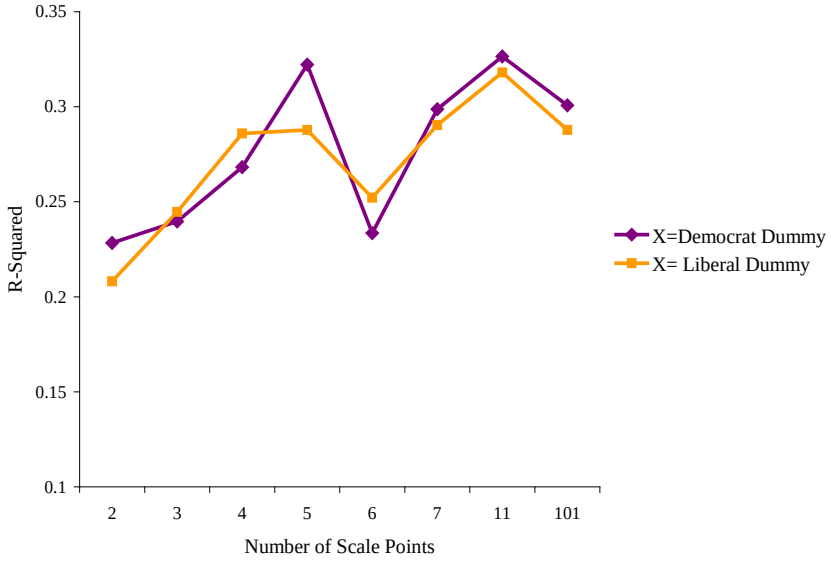


Figure 16: Estimated Effect of Party and Ideology on Partisan Attitudes, by Scale Length



Note: Entries are unstandardized regression coefficients obtained by regressing partisan attitudes—ratings of the 16 political targets, scaled to range from 0 (most liberal/Democratic) to 1 (most conservative/Republican—on a dummy variable for Democratic party identification and on a dummy variable for liberal self-identification, in turn. Analysis is based on the stacked data and included item fixed effects; results are essentially unchanged if we exclude item fixed effects.

Figure 17: Fit (R^2) from Regression of Partisan Attitudes on Party and Ideology, by Scale Length



Note: Entries are R-squared values from the regressions described in the note to Figure 16.

Appendix: Question Wording and Scale Presentation

Attitude Question Wording:

"We're **[also]** interested in your feelings towards a number of **[celebrities OR politicians OR groups]** in the news.

Please indicate **[whether OR the extent to which]** you feel favorable or unfavorable toward each of the people listed below."

[Note: Feeling Thermometers have additional instructions, printed below]

Celebrities:
Lindsay Lohan
Ben Affleck
Meryl Streep
Rosie O'Donnell
Jay-Z
Tina Fey
Justin Bieber
Clint Eastwood

Politicians:
John Boehner
Bill Clinton
George W. Bush
Joe Biden
Nancy Pelosi
Barack Obama
Paul Ryan
Mitt Romney

Groups:
Democrats
Gays & Lesbians
Labor Unions
Tea Party
Big Business
Republicans
Conservatives
Liberals

Political Knowledge Scale:

Do you happen to remember which party controls the United States House of Representatives – that is, which party has a majority of members in the United States House of Representatives?
[Answer Choices: Republicans, Democrats, I'm not sure]

Do you happen to remember which party controls the United States Senate – that is, which party has a majority of members in the United States Senate?
[Answer Choices: Republicans, Democrats, I'm not sure]

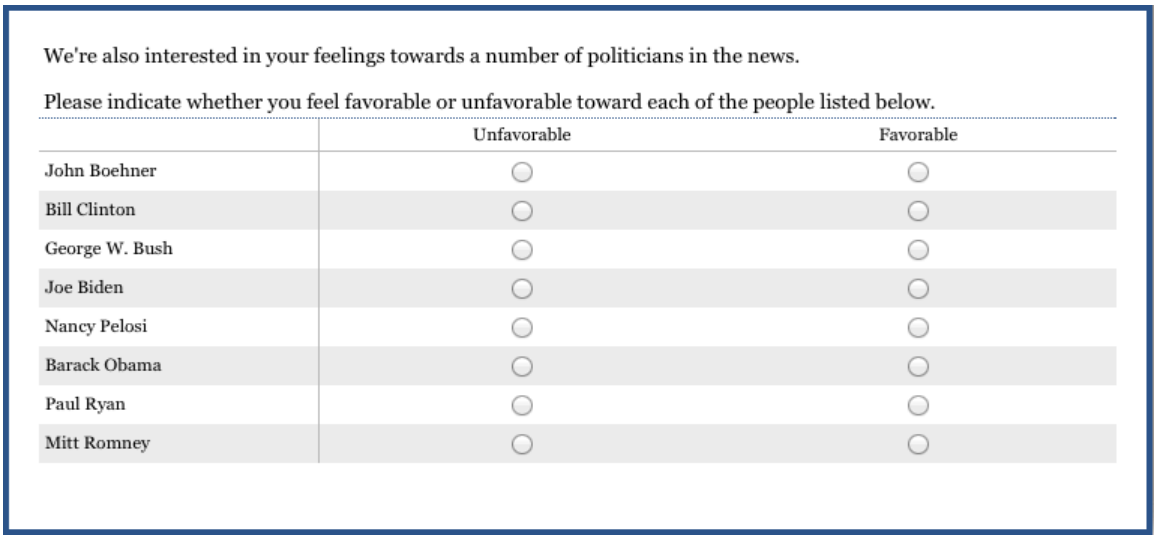
Do you happen to remember what job John Boehner holds?
[Answer Choices: Speaker of the US House, Governor of Texas, Chief Justice of the US Supreme Court, Prime Minister of Canada, Vice President of the United States, I'm not sure]

Do you happen to remember what job John Roberts holds?
[Answer Choices: Speaker of the US House, Governor of Texas, Chief Justice of the US Supreme Court, Prime Minister of Canada, Vice President of the United States, I'm not sure]

For how many years is a member of the United States Senate elected – that is, how many years are there in one full term of office for a US Senator?
[Answer Choices: Two years, Four years, Six years, Eight years, I'm not sure

Experimental Conditions:

2-point radial buttons:



We're also interested in your feelings towards a number of politicians in the news.

Please indicate whether you feel favorable or unfavorable toward each of the people listed below.

	Unfavorable	Favorable
John Boehner	○	○
Bill Clinton	○	○
George W. Bush	○	○
Joe Biden	○	○
Nancy Pelosi	○	○
Barack Obama	○	○
Paul Ryan	○	○
Mitt Romney	○	○

3-point radial buttons:

We're also interested in your feelings towards a number of politicians in the news.

Please indicate whether you feel favorable or unfavorable toward each of the people listed below.

	Unfavorable	Neither Unfavorable or Favorable	Favorable
John Boehner	○	○	○
Bill Clinton	○	○	○
George W. Bush	○	○	○
Joe Biden	○	○	○
Nancy Pelosi	○	○	○
Barack Obama	○	○	○
Paul Ryan	○	○	○
Mitt Romney	○	○	○

4-point radial buttons:

We're also interested in your feelings towards a number of politicians in the news.

Please indicate the extent to which you feel favorable or unfavorable toward each of the people listed below.

	Very Unfavorable	Unfavorable	Favorable	Very Favorable
John Boehner	○	○	○	○
Bill Clinton	○	○	○	○
George W. Bush	○	○	○	○
Joe Biden	○	○	○	○
Nancy Pelosi	○	○	○	○
Barack Obama	○	○	○	○
Paul Ryan	○	○	○	○
Mitt Romney	○	○	○	○

5-point radial buttons:

We're also interested in your feelings towards a number of politicians in the news.

Please indicate the extent to which you feel favorable or unfavorable toward each of the people listed below.

	Very Unfavorable	Unfavorable	Neither Unfavorable or Favorable	Favorable	Very Favorable
John Boehner	○	○	○	○	○
Bill Clinton	○	○	○	○	○
George W. Bush	○	○	○	○	○
Joe Biden	○	○	○	○	○
Nancy Pelosi	○	○	○	○	○
Barack Obama	○	○	○	○	○
Paul Ryan	○	○	○	○	○
Mitt Romney	○	○	○	○	○

6-point radial buttons:

We're also interested in your feelings towards a number of politicians in the news.

Please indicate the extent to which you feel favorable or unfavorable toward each of the people listed below.

	Very Unfavorable	Unfavorable	Slightly Unfavorable	Slightly Favorable	Favorable	Very Favorable
John Boehner	○	○	○	○	○	○
Bill Clinton	○	○	○	○	○	○
George W. Bush	○	○	○	○	○	○
Joe Biden	○	○	○	○	○	○
Nancy Pelosi	○	○	○	○	○	○
Barack Obama	○	○	○	○	○	○
Paul Ryan	○	○	○	○	○	○
Mitt Romney	○	○	○	○	○	○

7-point radial buttons:

We're also interested in your feelings towards a number of politicians in the news.

Please indicate the extent to which you feel favorable or unfavorable toward each of the people listed below.

	Very Unfavorable	Unfavorable	Slightly Unfavorable	Neither Unfavorable or Favorable	Slightly Favorable	Favorable	Very Favorable
John Boehner	○	○	○	○	○	○	○
Bill Clinton	○	○	○	○	○	○	○
George W. Bush	○	○	○	○	○	○	○
Joe Biden	○	○	○	○	○	○	○
Nancy Pelosi	○	○	○	○	○	○	○
Barack Obama	○	○	○	○	○	○	○
Paul Ryan	○	○	○	○	○	○	○
Mitt Romney	○	○	○	○	○	○	○

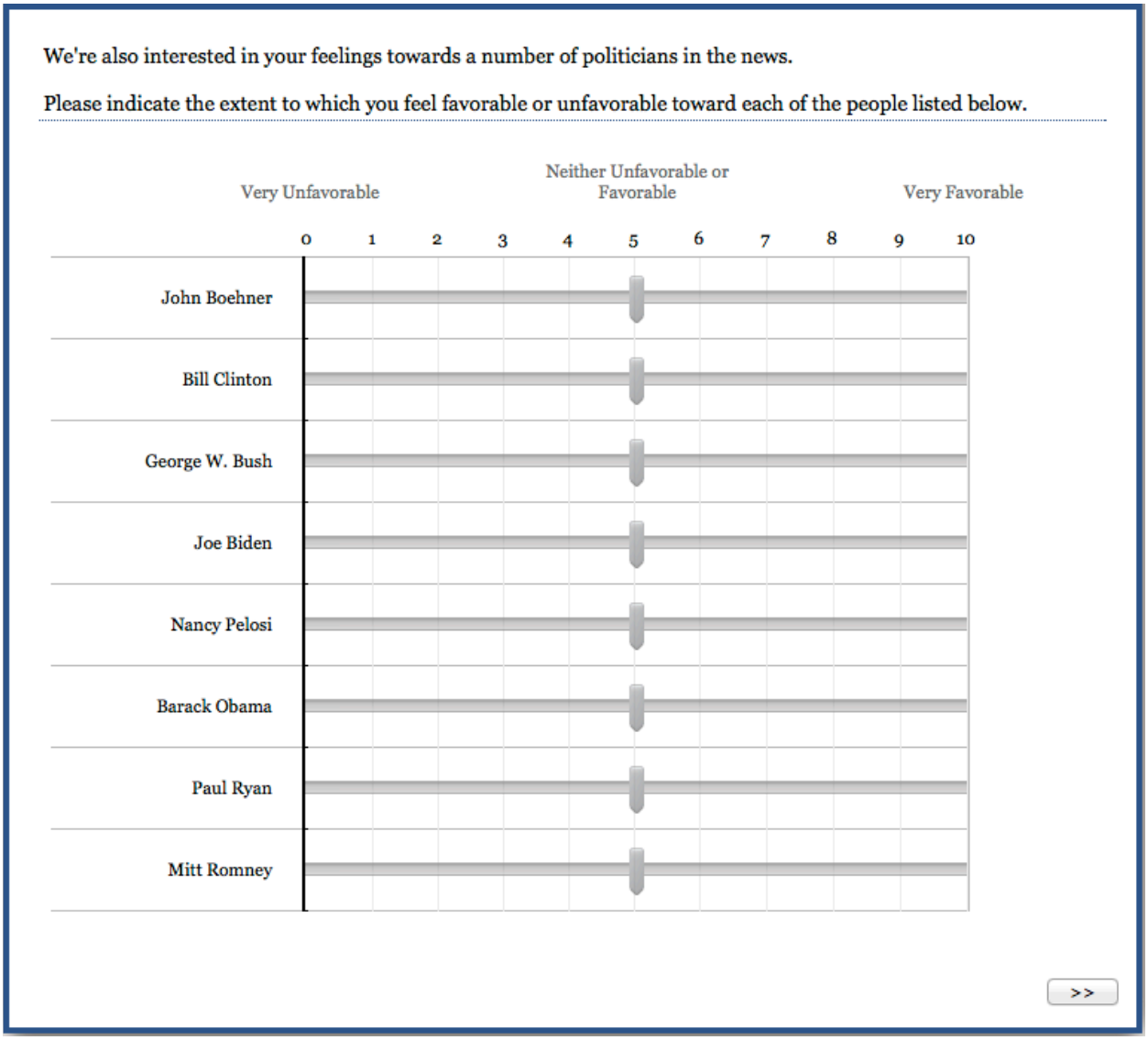
11-point radial buttons (with three labels):

We're also interested in your feelings towards a number of politicians in the news.

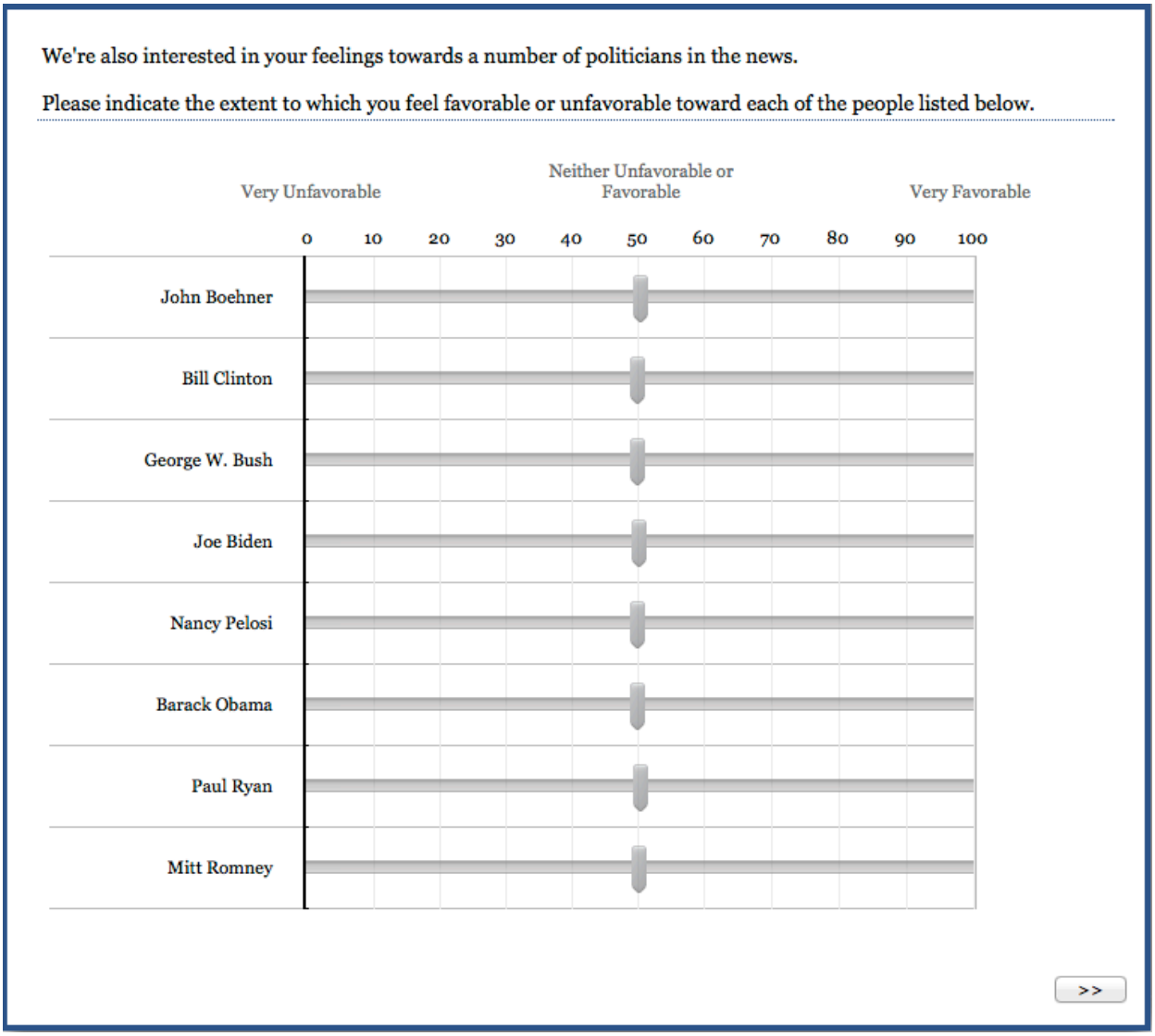
Please indicate the extent to which you feel favorable or unfavorable toward each of the people listed below.

	Very Unfavorable					Neither Unfavorable or Favorable					Very Favorable
John Boehner	○	○	○	○	○	○	○	○	○	○	○
Bill Clinton	○	○	○	○	○	○	○	○	○	○	○
George W. Bush	○	○	○	○	○	○	○	○	○	○	○
Joe Biden	○	○	○	○	○	○	○	○	○	○	○
Nancy Pelosi	○	○	○	○	○	○	○	○	○	○	○
Barack Obama	○	○	○	○	○	○	○	○	○	○	○
Paul Ryan	○	○	○	○	○	○	○	○	○	○	○
Mitt Romney	○	○	○	○	○	○	○	○	○	○	○

11-point slider scale (with three labels):



101-point slider scale (with three labels):



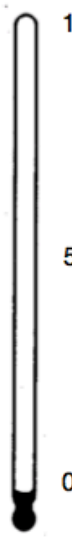
101-point slider scale (with seven labels):

101-point feeling thermometer (with three labels):

Please look at the graphic below.

We would like to get your feelings toward some politicians who are in the news these days. We will show the name of a person and we'd like you to rate that person using something we call the feeling thermometer.

Ratings between 50 degrees and 100 degrees mean that you feel favorable and warm toward the person. Ratings between 0 degrees and 50 degrees mean that you don't feel favorable toward the person and that you don't care too much for that person. You would rate the person at the 50 degree mark if you don't feel particularly warm or cold toward the person.



	Feeling Thermometer Rating
John Boehner	
Bill Clinton	
George W. Bush	
Joe Biden	
Nancy Pelosi	
Barack Obama	
Paul Ryan	
Mitt Romney	

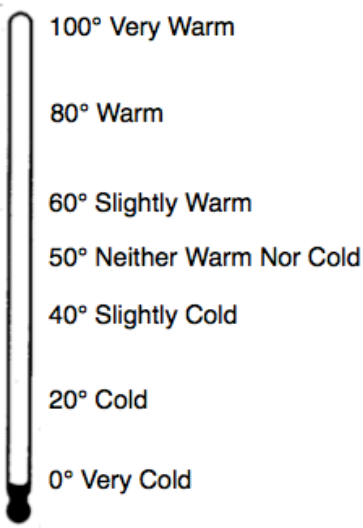
>>

101-point feeling thermometer (with seven labels):

Please look at the graphic below.

We would like to get your feelings toward some politicians who are in the news these days. We will show the name of a person and we'd like you to rate that person using something we call the feeling thermometer.

Ratings between 50 degrees and 100 degrees mean that you feel favorable and warm toward the person. Ratings between 0 degrees and 50 degrees mean that you don't feel favorable toward the person and that you don't care too much for that person. You would rate the person at the 50 degree mark if you don't feel particularly warm or cold toward the person.



	Feeling Thermometer Rating
John Boehner	
Bill Clinton	
George W. Bush	
Joe Biden	
Nancy Pelosi	
Barack Obama	
Paul Ryan	
Mitt Romney	

>>