Achieving Efficiency Without Losing Accuracy: Strategies for Scale Reduction with an Application to Risk Attitudes and Racial Resentment*

Krista Loose, *Government Accountability Office*
Yue Hou, *University of Pennsylvania*
Adam J. Berinsky, *Massachusetts Institute of Technology*

Objectives. Researchers often employ lengthy survey instruments to tap underlying phenomena of interest. However, concerns about the cost of fielding longer surveys and respondent fatigue can lead scholars to look for abbreviated, yet accurate, variations of longer, validated scales. In this article, we provide a template to aid in scale reduction. *Methods.* The template we develop walks researchers through a procedure for using existing data to consider all possible subscales along several reliability and validity criteria. We apply our method to two commonly used scales: the seven-item Risk Attitudes Scale and the six-item Racial Resentment Scale. *Results.* After applying the template, we find a four-item Risk Attitudes Scale that maintains nearly identical reliability and validity as the full scale and a three-item Racial Resentment Subscale that outperforms the two-item Subscale currently used in a major congressional survey. *Conclusions.* Our general template should be of use to a broad range of scholars seeking to achieve efficiency without losing accuracy when reducing lengthy scales. The code to implement our procedures is available as an R package, ScaleReduce.

Social scientists work with phenomena that are difficult to measure. The theories we develop often involve variables that are not directly observable, such as loyalty to a party, impressions of political candidates, predispositions toward political activity, or emotional reactions to unfolding events. Tests of theories involving such variables require measures that approximate these unobservable concepts. Typically, researchers use multiple survey responses to tap the underlying quantity of interest. For instance, risk attitudes (seven items), racial resentment (four items), moral traditionalism (six items), and egalitarianism (six items) are all multi-item measures commonly used in surveys employed by social scientists (Conover and Feldman, 1986; Feldman, 1988; Kam and Simas, 2010; Kinder and Sanders, 1996). Nevertheless, public opinion scholars face pressures in the form of limited survey space and the conventions of survey administration. Indeed, even scales of moderate length might strain a given researcher's resources (and a typical survey respondent's patience). However, to date, there has been little scientific work in social sciences to guide

*Direct correspondence to Krista Loose, Government Accountability Office, 10 Causeway Street #575, Boston, MA 02222 ⟨loosek@gao.gov⟩. The opinions and views expressed in this article are the authors' alone and are not intended to reflect GAO's institutional views. The authors thank Cindy Kam, who not only provided risk attitudes data for this project, but also helped them to aim higher in their goals for scale reduction and provided helpful feedback throughout. They also acknowledge the American National Election Study and the Cooperative Congressional Election Study for making their historic data on racial resentment available. Finally, the authors thank Neal Schmitt, who tirelessly read several drafts of this article, and Daniel Guenther and Robert Pressel for their help in editing.

SOCIAL SCIENCE QUARTERLY, Volume 99, Number 2, June 2018

DOI: 10.1111/ssqu.12414

scale reduction efforts. Here, we outline a principled strategy for identifying the most efficient reduced survey scales derived from more lengthy measures. Our approach for condensing multiple-item scales is applicable to any survey researcher who wishes to measure a variable of interest more compactly. We also provide an R package to aid researchers in implementing our suggested procedures.

To provide concrete examples, we apply this principled strategy to two scales: Risk Attitudes and Racial Resentment. Comprising seven items, the Risk Attitudes Scale (Kam, 2012; Kam and Simas, 2010, 2012) is relatively compact but may still be cost prohibitive—both in terms of respondent fatigue and survey administration time. Our method identifies a promising four-item subscale. The other scale, Racial Resentment, a commonly used measure of modern racism, was originally administered as a six-item scale (Kinder and Sanders, 1996) and has been reduced over time to a two-item scale in the Cooperative Congressional Election Study (CCES). In this case, our data analysis suggests that researchers have reduced the scale too far.

Condensing a Multiple-Item Scale

A core challenge for social scientists in developing measures is balancing the demands of accuracy and efficiency. In this article, we advocate the use of multi-item scales. Using a single item to tap an underlying personality dimension may be problematic, as any single item is prone to measurement error (Lord, Novick, and Birnbaum, 1968; Wiley and Wiley, 1971; Achen, 1975). Short measures can understate the importance of the underlying trait but also, by extension, overstate the importance of other variables. Moreover, reliability nearly always increases with multiple items. But how to select the number and type of those items is a complicated undertaking.

Given practical limitations, we want to measure our concept of interest with the fewest items possible while retaining the core benefits of the longer scale. Estimates from YouGov and GfK (formerly Knowledge Networks), two prominent Internet survey companies, suggest that for a nationally representative survey sample of 1,000 respondents, each additional question costs between $150 and $275 and takes, on average, 15 seconds to answer.

In psychology and other fields, abundant work has validated shortened scales. However, much of this work has relied on only one or two simple statistics or tests, such as Cronbach's alpha, item-scale correlations, or predictive validity (Stanton et al., 2002; Bizer et al., 2004; Zaller, 1990), to assert that the shorter scale comports with the original. Others have used additional criteria—such as similarity of responses across various demographic groups (Hoyle et al., 2002), item difficulty (Delli Carpini and Keeter, 1993), or diversity (Price and Zaller, 1993)—but few systematic approaches to scale reduction exist.[1]

[1]We have uncovered a few exceptions. Montgomery and Cutler (2013), who use computerized adaptive testing (CAT), where questions to measure an underlying trait are selected dynamically based on previous responses. There are some drawbacks to this approach, however, in that it may not be appropriate for all types of scales, is somewhat challenging to implement, and cannot be administered in a pen-and-paper format. Smith, McCarthy, and Anderson (2002) identify nine "sins" involved with shortening a scaled measure. Many of these sins involve insufficient consideration of the multiple content areas involved in many scales or inadequate a priori analyses of what might be gained or lost by shortening a scale. Second, Stanton et al. (2002) suggest researchers consider three aspects of the scale when making decisions about eliminating items: (1) internal item qualities, such as Cronbach's alpha, (2) external item qualities, such as the revised scale's correlation with other criteria, and (3) judgmental item qualities, including expert opinion about the breadth and importance of various questions. While Stanton and his co-authors provide some best practice guidelines for scale reduction,

We provide a template for evaluating both the length and item composition of a reduced scale in a more systematic and comprehensive way than typically employed in the literature. We conduct statistical analyses of all possible subscales of the original scale, using the performance of the full set of possible subscales to inform our decision making. We advocate that researchers take a broad view when considering available statistics that we know reflect various properties of scale validity and reliability. Our method can be easily applied to scales of any length and in any content area and we hope it will both broaden and systematize the criteria upon which future research approaches issues of scale reduction.[2]

Criteria for Evaluating Reduced Scales

The first criterion we discuss is *correlation* with the original scale. A reduced scale should provide measures of the latent concept that are quite similar to the full scale. To our knowledge, there are no established metrics for how closely shortened scales should correlate with the original, so evaluation here must be made relative to other possible shortened scales.

The second criterion is *cross-item reliability, or intercorrelation*. In the context of a multiple-item scale, the items should be highly intercorrelated if they are measuring the same underlying construct (see, e.g., Aronson et al., 1990). Intercorrelation between items is often measured in three ways:

- Cronbach's coefficient alpha (α), which estimates the degree to which each item in a scale is measuring the same underlying concept (DeVellis, 1991; Zeller and Carmines, 1979). Typically, researchers developing new scales aim for alpha coefficients of at least 0.70 (Schmitt, 1996). Alpha coefficients typically increase with scale length (Schmitt, 1996), however, which may complicate analyses of shorter scales.
- Item-scale correlations (also known as item-total), which identify how much each individual item correlates with the scale composed of all other items (DeVellis, 1991). An item that scores low on this measure is generally considered to be inconsistent with the other items that make up the proposed scale.
- Factor analysis, which determines the degree to which each item taps a single construct or is inherently multidimensional. Typically, researchers identify eigenvalues greater than 1 as indicating a dimension worthy of study (Fabrigar et al., 1999).

Our third criterion is *cross-time reliability*: If we assume that an underlying trait is stable across time, then reliable items (and scales constructed from such items) should correlate highly over time. Cross-time reliability requires administration of the proposed scale more than once to the same respondents.

they do not speak to questions of ideal scale length but simply assume a preferred length for a shortened scale. As we show in our racial resentment example, this approach could be problematic. Finally, Jacoby (1998) did some limited analyses on three scales, including racial resentment. He considered four criteria. Notably, Jacoby (1998) dismissed validity concerns about the reduced scales, indicating that such issues would be addressed by assuring the full scale was valid.

[2]Because we consider all possible subscales, this process could become cumbersome with very lengthy scales. In our longer case, the seven-item Risk Attitudes Scale produces 127 subscales. A 15-item scale, however, produces an exponentially larger 32,767 subscales. The code we provide will work with scales of any length, but note that the analysis of subscales based upon longer scales will take significant computing time. In our tests, the computing time approximately doubled for each additional item in the full scale. When working with longer scales, it may be more efficient to analyze specific facets separately rather than analyze all items for the entire scale (if applicable) or to randomly split the scale into smaller components and identify potentially problematic items.

Our fourth criterion is *classification*. Classification analysis considers whether the shortened scale and the original scale identify the same group of respondents. In the case of the Risk Attitudes Scale, for example, we can gauge whether the shorter scale distinguishes risk-accepting versus risk-averse individuals as well as the original scale does. For some scales, it is unclear what would be gained by reducing the variation of continuous measures.[3] Nonetheless, as an example drawn from the psychological literature, many scales are designed to classify patients as symptomatic for particular psychiatric disorders, such as schizophrenia. A well-performing reduced scale for schizophrenia symptoms should be designed classify the same individuals as the original longer form.

Scale reduction necessarily entails some tradeoffs. By improving a scale's reliability, we may compromise the validity of the measure. Specifically, we may no longer be measuring the full content of the underlying trait (Smith, McCarthy, and Anderson, 2002; Stanton et al., 2002). Therefore, our final criterion concerns *validity*. While we continue, in most cases, to perform our analyses on the full array of possible subscales, our validity analyses will focus on how a proposed shortened scale performs relative to the full scale. Specifically, we will examine two types of validity:

- *criterion validity*, or the extent to which the reduced scale correlates with demographic benchmarks from the literature, as well as other related measures with which it should theoretically correlate; and
- *predictive validity,* to examine the correlations between the reduced scale and a variety of political attitudes and participatory behaviors.

Approach to Analyzing Scales

Based on a review of the scale reduction literature in both political science and psychology, it might appear that these criteria are substitutable: researchers typically only use one or two to justify their conclusions. We argue, however, that the criteria are instead complementary. Analyzing measures of all criteria jointly will give researchers a better picture of the properties of both the full scale and potential reduced scales. In all cases, we encourage a transparent and systematic process for selecting a reduced scale. Researchers should be explicit about the criteria considered and any results that may potentially favor a different subscale than the one selected. Evaluating a more expansive array of statistics will lead to a more informed choice of subscale. However, this approach risks having different criteria pointing to different subscales. For example, some subscales might perform better in predictive validity, while others might have superior criterion validity. In that case, we offer the following advice for best interpreting the findings and choosing the right subscale:

- First, consider whether some metrics may be artificially inflated. Shared question format and/or response options may encourage similar responses and result in high α statistics that are artificially inflated as an artifact of the common instrumentation. For instance, acquiescence bias may artificially inflate α if the items all point in the same direction.
- Second, determine which metrics are most important for their research. For example, it may be less common for political scientists, compared with, say, clinical psychologists,

[3] Some works have constructed risk preference as a binary variable. For instance, Berinsky and Lewis (2007) construct a scale of risk proclivities by splitting their sample at the median into two groups: high-risk and low-risk takers.

to need to accurately classify respondents into "types," so this criterion may be less critical.
- Third, consider whether inconsistent results, should they emerge, reveal anything about the properties of the full scale. Conflicting results may indicate underlying problems with items within the full scale or the existence of a scale that may not be amenable to reduction.
- Finally, evaluate whether changes to the questions themselves would be beneficial. Sometimes, condensing long scales may require researchers to modify existing items, generate new items, and response scales, or construct item-specific response options (Bizer et al., 2004).

Our key contributions in this approach are threefold. First, we show that using the full set of possible subscales enriches the decision-making process. Second, we propose a five-tiered approach to assessing performance. Third, we provide a systematic method for triangulating on an accurate, but more efficient, reduced scale.

Empirical Examples

We now present two examples of reducing a scale using our method. In the first, we examine the seven-item Risk Attitudes Scale and find a four-item reduced scale that performs nearly identically to the full scale. Table 1 lists the seven questions that make up the original scale. Throughout, we will refer to specific questions using the mnemonic provided in the first column of the table. The seven-item battery has been fielded many times over the past several years; Supporting Information Appendix A provides details on data we used in this article.[4]

The second example we evaluate is the commonly used Racial Resentment Scale. Initially developed as a six-item scale and fielded as such in the 1986 American National Election Study (ANES), by the 1988 ANES, it was reduced to a four-item scale. More recently, the CCES has begun using two of the questions as an abbreviated scale. Table 2 contains the original six items, along with the item mnemonic. Table 2 also notes whether each question is contained in the shortened four-item scale and in the CCES two-item scale. This scale has also been administered many times; Supporting Information Appendix B provides additional information on the particular data sets we use here.

In each of the examples, we begin with a brief summary of how the scale has been used in the literature. We then generate all possible combinations of the original scales. This yields $\sum_{k=1}^{7}\binom{7}{k}=127$ possible combinations in the case of risk attitudes and $\sum_{k=1}^{6}\binom{6}{k}=63$ combinations of racial resentment items. Where possible, we conduct all analyses on each identified subscale.

The analyses are produced using the freely available R package, ScaleReduce, and are summarized in Tables 3–6, two tables for each scale. Tables 3 (risk attitudes) and 4 (racial resentment) present the statistical properties of scales of varying lengths. The data are aligned by row, starting with the one-item scales and ending with the single seven- (six)-item scale. The number of items in each scale is indicated in the first column, and the

[4]Smith, McCarthy, and Anderson (2002) and Stanton et al. (2002) encourage researchers to administer the shortened scale on its own, as they note that responses may differ when separated from the context of the larger scale. We do not do that for this article, but we do note that one of our administrations asked the risk battery in a different order, which gives us confidence that our results are not driven by question order effects.

TABLE 1

Risk Attitudes Items

Item Mnemonic	Question Wording Response Options
Cautious	Some people say you should be cautious about making major changes in life. Suppose that these people are located at 1. Others say that you will never achieve much in life unless you act boldly. Suppose these people are located at 7. And others have views in between. Where would you place yourself on this scale? *1 = "One should be cautious about making major changes in life" to 7 = "One will never achieve much in life unless one acts boldly"*
Horse	Suppose you were betting on horses and were a big winner in the third or fourth race. Would you be more likely to continue playing or take your winnings? *Definitely continue playing, probably continue playing, not sure, probably take my winnings, definitely take my winnings*
Risks	In general, how easy or difficult is it for you to accept taking risks? *Very easy, somewhat easy, somewhat difficult, very difficult*
	Please rate your level of agreement or disagreement with the following statements. *Strongly disagree, disagree, neither disagree nor agree, agree, strongly agree*
Explore	I would like to explore strange places.
Frighten	I like to do frightening things.
Experiences	I like new and exciting experiences, even if I have to break the rules.
Friends	I prefer friends who are exciting and unpredictable.

NOTE: Items identically worded in all data sets.

number of instances a subscale of that length is included in our data is indicated in the second column. The tables also provide information on a credible interval of estimates: the range containing the middle 95 percent of estimates for all subscales of that length or containing that item.[5]

We also want to assess the relative strengths of each specific question. The data in Tables 5 (risk attitudes) and 6 (racial resentment), therefore, examine the statistical properties of scales containing particular questions. The data are aligned by row, marked by each specific question. After we summarize the results in these tables, we provide additional tests of subscale validity.

Risk Attitudes

The seven-item Risk Attitudes Scale developed by Kam and Simas (2010, 2012; Kam, 2012) builds a bridge between the long survey batteries favored by psychology[6] and several

[5]Although we provide 95 percent credible intervals, we note that particular instances of subscales may have outlying values that are worth noting.

[6]Psychologists have developed quite lengthy measures to tap individual differences in risk attitudes. Zuckerman's Sensation-Seeking Scale (Zuckerman, Eysenck, and Eysenck, 1978) consists of 40 items that tap thrill

TABLE 2

Racial Resentment Items

Item Mnemonic	ANES Four-Item	CCES Two-Item	Question Wording Response Options
			Do you agree with the following statements? *Strongly disagree, disagree, neither disagree nor agree, agree, strongly agree*
Work up	Yes	Yes	Irish, Italians, Jewish, and many other minorities overcame prejudice and worked their way up. Blacks should do the same without any special favors.
Slavery	Yes	Yes	Generations of slavery and discrimination have created conditions that make it difficult for blacks to work their way out of the lower class."
Deserve	Yes	No	Over the past few years, blacks have gotten less than they deserve.
Try harder	Yes	No	It is really a matter of some people not trying hard enough; if blacks would only try harder they could be just as well off as whites.
Welfare	No	No	Most blacks who receive money from welfare programs could get along without it if they tried.
Attention	No	No	Government officials usually pay less attention to a request or complaint from a black person than from a white person.

NOTE: Items identically worded in all data sets.

one-item approaches used in political science.[7] The scale features individual questions that have been used as valid indicators of risk attitudes across several disciplines. The questions vary in direction in order to avoid acquiescence response bias. However, there may be room for additional efficiency gains in measuring risk attitudes. Tables 3 and 5 offer preliminary evidence on the minimum number of items that should be included in an efficient yet well-performing risk attitudes subscale as well as the specific items that might be included therein.

These exercises suggest that a four-item scale may be appropriate. A four-item scale is likely to correlate at above 0.90 with the full scale and a factor analysis of those items is also likely to return a first eigenvalue greater than 1. The test–retest correlations for four-item scales are also quite high, with such correlations dropping more dramatically as the scale drops in length. A four-item scale also performs well at classifying individuals as risk tolerant or risk averse. Although the average Cronbach's α for four-item scales falls below the typically used 0.70 threshold (Schmitt, 1996), four-item scales that do *not* include the *horse* item all have α values greater than 0.68. This result illustrates the value

and adventure seeking, experience seeking, disinhibition, and boredom susceptibility. Another instrument for assessing risk orientations is the Choice Dilemmas Questionnaire (CDQ) that consists of 12 separate vignettes, in which subjects are asked to provide advice to an individual facing some uncertain situation. Finally, the 30-item Domain-Specific Risk Taking Scale (DOSPERT) identifies risky behaviors across six dimensions, including social, recreational, financial, economic, health, and ethical risks (Blais and Weber, 2006).

[7]Some political scientists have found that risk orientations affect trust in government, vote choice, and opinions about public policies (Peterson and Lawson, 1989; Morgenstern and Zechmeister, 2001; Ehrlich and Maestas, 2010); others have found that risk orientations predict willingness to tolerate ambiguous candidate positions (Tomz and Van Houweling, 2009).

TABLE 3

Summary Statistics of Subscales— Risk Attitudes, by Number of Items

No. of Items	No. of Scales	Correlation with Full Scale	Cronbach's α	First Eigenvalue	Second Eigenvalue	Test–Retest Correlation	Proportion Correctly Classified
1	7	**0.64** [0.45, 0.72]				**0.63** [0.50, 0.70]	**0.71** [0.64, 0.79]
2	21	**0.79** [0.68, 0.84]	**0.46** [0.18, 0.63]	**0.42** [0.09, 0.70]	**−0.20** [-0.25, -0.08]	**0.69** [0.63, 0.75]	**0.78** [0.72, 0.83]
3	35	**0.87** [0.82, 0.90]	**0.56** [0.35, 0.70]	**0.83** [0.38, 1.21]	**−0.08** [-0.16, 0.00]	**0.73** [0.68, 0.78]	**0.82** [0.77, 0.86]
4	35	**0.92** [0.90, 0.93]	**0.63** [0.51, 0.75]	**1.21** [0.82, 1.62]	**−0.02** [-0.09, 0.03]	**0.76** [0.73, 0.80]	**0.86** [0.82, 0.88]
5	21	**0.95** [0.95, 0.96]	**0.68** [0.62, 0.77]	**1.58** [1.31, 1.99]	**0.02** [-0.02, 0.07]	**0.78** [0.76, 0.81]	**0.89** [0.87, 0.92]
6	7	**0.98** [0.97, 0.99]	**0.72** [0.70, 0.78]	**1.94** [1.78, 2.25]	**0.05** [0.01, 0.09]	**0.80** [0.78, 0.81]	**0.92** [0.90, 0.95]
7	1		**0.75**	**2.30**	**0.09**	**0.81**	

NOTE: Table entry is the **mean** estimate with the 95 percent credible interval in brackets—the first value is the estimate at the 2.5th quantile, the second value is the estimate at the 97.5th quantile. For instance, the third row documents relevant summary statistics of all 35 subscales that contain three items. The "correlation with full scale" entry documents the correlations between all 35 subscales and the full seven-item scale: 0.87 is the mean value, 0.82 is the 2.5th quantile value, and 0.90 is the 97.5th quantile value.
SOURCE: KN2008, except for test–retest analysis, which used the 2009 Knowledge Networks panel data.

TABLE 4

Summary Statistics of Subscales—Racial Resentment, by Number of Items

No. of Items	No. of Scales	Correlation with Full Scale	Cronbach's α	First Eigenvalue	Second Eigenvalue	Test-Retest Correlation*	Proportion Correctly Classified
1	6	**0.68** [0.53, 0.75]				**0.53** [0.49, 0.56]	**0.73** [0.67, 0.77]
2	15	**0.83** [0.74, 0.87]	**0.52** [0.28, 0.71]	**0.51** [0.19, 0.87]	**−0.22** [−0.25, −0.14]	**0.61** [0.60, 0.64]	**0.81** [0.76, 0.85]
3	20	**0.90** [0.86, 0.92]	**0.63** [0.51, 0.75]	**0.98** [0.64, 1.36]	**−0.08** [−0.15, −0.01]	**0.65** [0.65, 0.66]	**0.85** [0.82, 0.89]
4	15	**0.95** [0.93, 0.96]	**0.69** [0.64, 0.77]	**1.41** [1.18, 1.75]	**0.04** [−0.06, 0.21]	**0.68**	**0.89** [0.86, 0.91]
5	6	**0.98** [0.97, 0.98]	**0.74** [0.72, 0.78]	**1.82** [1.67, 2.06]	**0.17** [0.07, 0.26]		**0.93** [0.91, 0.95]
6	1		**0.77**	**2.23**	**0.29**		

NOTE: Table entry is the **mean** estimate with the 95 percent credible interval in brackets—the first value is the estimate at the 2.5th quantile, the second value is the estimate at the 97.5th quantile. For instance, the third row documents relevant summary statistics of all four subscales that contain three items. The "correlation with full scale" entry documents the correlations between all four subscales and the full four-item scale: 0.90 is the mean value, 0.86 is the 2.5th quantile value, and 0.92 is the 97.5th quantile value.

*The only panel administrations of the Racial Resentment Scale were done on the four-item scale.

SOURCE: ANES 1986, except for test–retest analysis, which used the 1990–1992 ANES panel data.

TABLE 5

Summary Statistics of Subscales—Risk Attitudes, by Question

Question	No. of Scales	Correlation with Full Scale	Cronbach's α	Item-Scale Correlations	First Eigenvalue	Second Eigenvalue	Test–Retest Correlation	Proportion Correctly Classified
Cautious	63	**0.90** [0.72, 0.98]	**0.60** [0.35, 0.76]	**0.36** [0.22, 0.46]	**1.17** [0.31, 2.08]	**−0.02** [−0.22, 0.08]	**0.74** [0.63, 0.80]	**0.85** [0.74, 0.92]
Horse	63	**0.89** [0.68, 0.98]	**0.54** [0.22, 0.73]	**0.17** [0.12, 0.19]	**1.06** [0.14, 1.99]	**−0.01** [−0.12, 0.07]	**0.75** [0.66, 0.81]	**0.85** [0.72, 0.93]
Risks	63	**0.91** [0.76, 0.98]	**0.65** [0.46, 0.77]	**0.52** [0.38, 0.61]	**1.30** [0.56, 2.13]	**−0.04** [−0.24, 0.08]	**0.76** [0.66, 0.81]	**0.86** [0.79, 0.93]
Explore	63	**0.91** [0.76, 0.98]	**0.63** [0.41, 0.77]	**0.46** [0.31, 0.55]	**1.25** [0.48, 2.13]	**−0.04** [−0.24, 0.08]	**0.75** [0.65, 0.81]	**0.85** [0.74, 0.93]
Frighten	63	**0.91** [0.78, 0.98]	**0.64** [0.39, 0.77]	**0.48** [0.26, 0.58]	**1.27** [0.36, 2.13]	**−0.04** [−0.25, 0.08]	**0.76** [0.69, 0.81]	**0.86** [0.77, 0.93]
Experiences	63	**0.91** [0.77, 0.98]	**0.63** [0.38, 0.77]	**0.46** [0.25, 0.56]	**1.26** [0.34, 2.13]	**−0.04** [−0.25, 0.08]	**0.77** [0.71, 0.81]	**0.86** [0.78, 0.93]
Friends	63	**0.90** [0.73, 0.98]	**0.62** [0.38, 0.76]	**0.42** [0.28, 0.51]	**1.22** [0.40, 2.13]	**−0.04** [−0.24, 0.08]	**0.75** [0.66,0.81]	**0.85** [0.74, 0.92]

NOTE: Table entry is the **mean** estimate with the 95 percent credible interval in brackets—the first value is the estimate at the 2.5th quantile, the second value is the estimate at the 97.5th quantile. For instance, the first row documents relevant summary statistics of all 63 subscales that contain the *cautious* item. The "correlation with full scale" entry documents the correlations between these subscales and the full seven-item scale: 0.90 is the mean value, 0.72 is the 2.5th quantile value, and 0.98 is the 97.5th quantile value. Item-scale correlations correlate each item with the 63 subscales that do not contain it. The factor analysis omits subscales with only one item. All entries exclude the full scale, which by definition includes all seven items.
SOURCE: KN2008, except for test–retest analysis, which used the 2009 Knowledge Networks panel data.

TABLE 6

Summary Statistics of Subscales—Racial Resentment, by Question

Question	No. of Scales	Correlation with Full Scale	Cronbach's α	Item-Scale Correlations	First Eigenvalue	Second Eigenvalue	Test–Retest Correlation*	Proportion Correctly Classified
Deserve	32	**0.91** [0.72, 0.98]	**0.67** [0.54, 0.78]	**0.51** [0.38, 0.58]	**1.24** [0.52, 2.11]	**−0.01** [−0.24, 0.27]	**0.62** [0.52, 0.68]	**0.87** [0.75, 0.94]
Try	32	**0.92** [0.81, 0.98]	**0.67** [0.45, 0.78]	**0.51** [0.27, 0.63]	**1.27** [0.43, 2.11]	**−0.01** [−0.25, 0.27]	**0.63** [0.56, 0.68]	**0.86** [0.76, 0.94]
Workway	32	**0.92** [0.81, 0.98]	**0.67** [0.49, 0.78]	**0.53** [0.33, 0.64]	**1.29** [0.45, 2.11]	**−0.02** [−0.25, 0.27]	**0.64** [0.57, 0.68]	**0.87** [0.77, 0.94]
Slavery	32	**0.91** [0.74, 0.98]	**0.65** [0.46, 0.78]	**0.42** [0.30, 0.49]	**1.18** [0.39, 2.11]	**−0.03** [−0.24, 0.25]	**0.62** [0.51, 0.68]	**0.87** [0.78, 0.94]
Welfare	32	**0.91** [0.78, 0.98]	**0.66** [0.44, 0.78]	**0.48** [0.29, 0.58]	**1.24** [0.37, 2.11]	**−0.02** [−0.25, 0.27]		**0.87** [0.78, 0.94]
Attention	32	**0.90** [0.68, 0.98]	**0.60** [0.30, 0.76]	**0.29** [0.18, 0.39]	**1.10** [0.21, 1.99]	**0.02** [−0.20, 0.27]		**0.86** [0.73, 0.94]

NOTE: Table entry is the **mean** estimate with the 95 percent credible interval in brackets—the first value is the estimate at the 2.5th quantile, the second value is the estimate at the 97.5th quantile. For instance, the first row documents relevant summary statistics of all 32 subscales that contain the *deserve* item. The "correlation with full scale" entry documents the correlations between these subscales and the full four-item scale: 0.91 is the mean value, 0.72 is the 2.5th quantile value, and 0.98 is the 97.5th quantile value. Item-scale correlations correlate each item with the 32 subscales that do not contain it. The factor analysis omits subscales with only one item. All entries exclude the full scale, which by definition includes all four items.

*The only panel administrations of the Racial Resentment Scale were done on the four-item scale; these statistics are based on eight subscales containing each item.

SOURCE: ANES 1986, except for test–retest analysis, which used the 1990–1992 ANES panel data.

of using multiple measures for scale reduction. If we relied solely on Cronbach's alpha and the generally recognized 0.70 threshold, we might conclude that only six-item scales were as reliable as the full scale—a reduction that saves relatively little time or money. Looking at the evidence more broadly, however, we see that a four-item scale may indeed perform similarly to the full scale.

In the interests of efficiency, we considered specifically whether a three-item scale would work as well as a four-item scale. However, our analyses suggest that a three-item scale is too short: the first eigenvalue is unlikely to be greater than 1; the Cronbach's alpha of three-item scales is almost always much smaller than the full scale; and the test–retest correlations are statistically different from the full scale for three-quarters of the three-item scales. Moreover, we test all three-item scales on the predictive validity measures described below and none of these scales consistently returned coefficients that are similar to the coefficients we estimate when we use the full scale.

On the other hand, it is not clear that a five-item scale provides any substantial gains. Although all the five-item scales are more highly correlated with the full scale than any of the four-item scales, the shorter four-item scales are still very highly correlated with the full scale. Moreover, an equal percentage of four- and five-item subscales have alpha statistics that are statistically similar to the full scale. Additionally, there are many four-item scales that show test–retest correlations as robust as the five-item scales. In sum, although not all four-item scales perform as well as a five-item scale, there are enough well-performing four-item scales to suggest that adding a fifth item may not be worth the added cost.

With a four-item scale in mind, which three questions should we consider excluding? The evidence in Table 5 suggests candidates for exclusion are: *cautious, horse,* and *friends*.[8] These items were consistently low-performing across all of our statistical tests, especially the *horse* question. We again note that the differences are slight in some cases, but given the consistency of the findings, we believe the optimal four-item scale would include *explore*, *experiences, frighten,* and *risks*. This particular subscale correlates with the seven-item scale at 0.92, has a Cronbach's α of 0.75, has a test–retest correlation of 0.80, and correctly classifies more than 88 percent of respondents. Histograms comparing our proposed reduced scale, the full scale, and the other subscales on several dimensions appear in Supporting Information Appendix A.

We next turn to the issue of validity. As we noted above, scale reduction entails some tradeoffs: by improving a scale's reliability, we may compromise the validity of the measure. Specifically, we may no longer be measuring the full content of the underlying trait (Smith, McCarthy, and Anderson, 2002; Stanton et al., 2002). In the above analyses, we saw that the *horse* item does not scale well with the other questions. We could be concerned that by deleting it, we are not capturing the full range of risk attitudes. It is therefore crucial to evaluate whether the reduced scale has similar validity properties as the original seven-item scale. Our analyses show that the removal of the *horse* item (along with *cautious* and *friends*) either improves our validity or does not affect the results.

One measure of the validity of the abbreviated scale is whether it correlates with characteristics with which it ought to correlate, based on existing literature. We generate pairwise

[8]In some of the other seven data sets, some analyses show that the *friends* item performs better than *explore*. However, the differences are often subtle and the four-item scales comprised *experiences, frighten, risks,* and either of these two items have similar reliability and validity properties. If there are theoretical reasons to prefer the *explore* question over the *friends* question, researchers could substitute the *explore* question without concerns over significantly reduced reliability or validity. Full results for the scaling analyses in other data sets appear in the Supporting Information Appendix A.

correlations between the full scale and a variety of variables, such as gender, age, or other measures of risk. We also generate these correlations with all 126 possible subscales of the risk measure. Across all demographic, political, and psychological variables, we find similar relationships using the full or reduced scale. The one possible exception is a stronger correlation between youth and the reduced scale.[9] The 95 percent credible interval for these correlations is visualized in Supporting Information Fig. A1.

We close this example with two tests of predictive validity, in which we relate risk attitudes to two different outcomes: political participation and support for probabilistic policy options. We replicate published findings (Kam, 2012; Kam and Simas, 2010) on the relationship between risk attitudes and these outcomes using the full scale. We then compare the predictive validity of the reduced scale with the original findings.

First, we compare the relationship between risk attitudes and prospective and retrospective political participation, when risk attitudes are measured with the full scale or with our proposed reduced scale.[10] The relationship between the participation variable and the risk scales is quite similar across the two measures. Across all nine participation-related variables, both prospective and retrospective, the direction, magnitude, and statistical significance of the coefficients in the reduced-scale models are extremely similar to those in the original models with the full seven-item scale, with largely overlapping confidence intervals.[11] These results are shown graphically in the Supporting Information Figs. A2 and A3.

Second, we replicate the results on the relationship between risk acceptance and preferences for probabilistic policy options from Kam and Simas (2010).[12] Kam and Simas show that risk preferences are related to policy preferences for certain (as opposed to probabilistic) outcomes, using a version of the Asian disease problem first developed by Tversky and Kahneman (1981). The overall are results shown in Supporting Information Fig. A4, and suggest coefficients of very similar direction, magnitude, and statistical significance in all analyses, where again the confidence intervals substantially overlap.[13] Risk attitudes, measured by the seven-item scale or the reduced four-item scale, predict a preference for a probabilistic outcome over a certain outcome, regardless of how the outcome is framed.[14]

[9]While the difference in correlation between the two scales is not large, we probe the distinction. We find that although most of the items in the risk scale are positively correlated with youthfulness, as we would expect, the cautious question has no correlation for youthfulness and the horse question is actually negatively correlated with youthfulness. Both patterns highlight potential issues of validity for these questions, and they suggest we may be better served by excluding them from the scale, a suggestion that comports with the findings reported above.

[10]This analysis replicates Tables 3 and 4 in Kam (2012). Following the missing data convention in Kam (2012), we use six- and five-item scale values to substitute the seven-item scale values when missingness occurs. Similarly, we use three-item scale values to substitute the reduced four-item scale values when missingness occurs. There is little missingness in the data: only 3 percent of the subjects left one question unanswered and only 1 percent of the subjects left two questions unanswered.

[11]To test the equivalence of coefficients, we performed Vuong's closeness tests (1989) comparing a model including the full scale with a model including the reduced scale. We fail to reject the hypothesis that the given two models are equally close to the true data generating process for each of the 18 comparisons. It indicates that there is no difference in using either scale to predict each and every participation variables.

[12]This analysis replicates Table 2 in Kam and Simas (2010). We use the same strategy as we employ above to account for missing data. As above, there is very little missingness.

[13]We again perform Vuong's closeness test to compare pairs of models with the full scale to those with the reduced scale. We fail to reject the null hypothesis that the given two models are equally close to the true model for each of the three paired comparisons.

[14]It should be noted that Ehrlich and Maestas (2010) have argued for the use of a single-item measure of risk preferences, showing that it can perform similarly to longer scales. For scholars who want to use such

Racial Resentment

Racial attitudes are another area that has stimulated much attention in political science. A huge body of literature suggests that racism is the driving force behind white Americans' opposition to policies designed to help blacks, and has the strongest predictive power on "white opinion on issues of race" (Kinder and Mendelberg, 2000). Among a variety of measurements on racism, the Racial Resentment Scale developed by Kinder and Sanders (1990, 1996)—which builds upon earlier work by Kinder and Sears (1981) and McConahay (1986)—is the most commonly used in large-scale surveys including the ANES and more recently the CCES.[15]

Although Kinder and Sanders's (1996) original scale contained six items tapping racial resentment, the scale was almost immediately reduced by the ANES to four items. Kinder and Sanders note that the questions *welfare* and *attention* were eliminated because they explicitly invoke government or governmental policies. The four-item Racial Resentment Scale has been administered countless times since 1986, including almost every administration of the ANES.[16] More recently, the CCES has used a two-item subset of the Racial Resentment Scale in its biennial surveys. The CCES uses the *slavery* and *work up* items; its documentation provides no rationale for selecting a two-item scale or for selecting these particular two items. Our full approach can provide a more reasoned decision about the appropriate length and content of a reduced Racial Resentment Scale. Therefore, we conduct our analyses on the original 1986 ANES data to show that the two-item CCES scale is bested by a three-item scale consisting of *deserve*, *slavery*, and *work up*. As a precaution against item performance changing over the intervening decades, we note that the results are largely the same across the ANES time series. These results are shown in Supporting Information Appendix B.

Table 4 offers evidence regarding the minimum number of items that should be included in the subscale for racial resentment. Our analysis suggests that the best-performing subscale of the four-scale Racial Resentment Scale is a three-item subscale, instead of the two-item scale used in CCES. We first rule out one-item subscales, since they have a significantly lower correlation with the full scale, and a much lower classification rate compared to the two- and three-item subscales.

Two- and three-item subscales perform similarly in terms of correlation with full scale and classification rate. At first glance, it may appear that there are two-item scales that perform similarly well to three-item options. For example, the upper bound on the Cronbach's α for a two-item scale exceeds that of the three-item scale. However, this illustrates a challenge with scale reduction that we mention above but is worth emphasizing here. The four-item Racial Resentment Scale includes two items where the "agree" option indicates sympathy toward blacks and two items where the "agree" option signifies a lack of sympathy. When

scales, our method can provide some guidance about one-item scales by evaluating correlation with the full scale, test–retest correlation, classification, and reliability, though not on the measures of intercorrelation.

[15]This scale is not without its critics. For example, Gomez and Wilson (2006) note similar measures such as the Symbolic Racism Scale can suffer from attribution bias (see also Feldman and Huddy 2005; Sniderman et al. 1996). We selected the Racial Resentment Scale for its frequent use and strive only to show that a subscale provides results that are equivalent to the full scale.

[16]In 1998, the ANES used a two-item scale consisting of *deserve* and *work up*. According to an analysis by Jacoby (1998), this selection was appropriate given that it had the highest Cronbach's α value of all two-item scale options, had a high correlation with the original six-item scale, and the distribution of racial resentment was similar if one considered either the four- or two-item scales. In our data, this scale outperforms the CCES two-item scale on all metrics except classification. It also has larger Cronbach's α than about a third of the three-item scales, but is less strong than the three-item scales on other metrics, such as correlation with the original scale or classification.

response options are similarly coded, α statistics are artificially inflated (Podsakoff et al., 2003). In this case, the subscales that contain only similarly coded response options have α statistics more than 0.2 higher than those subscales that contain differently coded items. Once we account for this issue, it quickly becomes clear that three-item scales outperform two-item options.

In particular, there are three-item scales that reach the generally accepted benchmark for Cronbach's α of 0.70 and all generate a first eigenvalue over 1 in factor analysis. While the differences are less stark for correlation with the full scale and classification, it is clear that all three-item scales are more highly correlated and superior at classification than any two-item scale.

If we assume a three-item scale is preferred over a two-item scale, Table 6 will help adjudicate which item should be dropped. We see that *slavery* performs somewhat poorly on many of the metrics. However, subscales containing *try harder* have much smaller lower bounds than other questions and the *deserve* question performs equally poorly as *slavery* on some metrics. Unlike the Risk Attitudes Scale, it is not immediately clear from Table 6 that there are one or more items that consistently perform worse than the others. We must make a judgment call as to which three-item subset would be preferable.

We can use our analyses of validity to choose which three-item subset is preferable. In this case, as with risk attitudes, we look at a series of correlations with known benchmarks (criterion validity) and regressions on political attitudes (predictive validity). Specifically, we look at the full scale in relation to the four-item ANES scale, the two-item CCES scale, and all three-item subscales. For these analyses, we use as our guide Kinder and Sanders's (1996) description and testing of the original Racial Resentment Scale.

To begin with the correlations with known benchmarks, we evaluate how racial resentment correlates with a variety of measures Kinder and Sanders (1996) identified as related to their new measurement, such as ideology, race of interviewer, and perception of black stereotypes. There are somewhat limited differences between the full scale and these six subscales.

Overall, these analyses do not give us confidence in choosing a particular three-item subscale. For the most part, the three-item subscales seem to perform similarly across these measures of criterion validity. The only scale we might potentially want to remove from consideration is the three-item scale that does not contain *slavery*, as this scale was noticeably different on correlations with race of interviewer and biological racism. These relationships are illustrated in Supporting Information Fig. B1.

As a final analysis, we evaluate how well the subscales of racial resentment predict attitudes toward explicitly racial government policies. We replicate the analyses in Table 5.5 in Kinder and Sanders (1996:117). As we note above with risk attitudes, we are interested in whether particular subscales are able to identify *the same relationship* as does the full four-item scale.

In the models where we use the two-item scale to predict government policies, the coefficients are significantly lower than those estimated using the original scale. There is much more consistency here in terms of which three-item scale is preferred, with the three-item scale that drops the *try harder* item generally providing coefficients quite similar to the four-item scale.[17] These results are shown in Supporting Information Fig. B2.

[17]Across all policy areas, the six-item scale produces the strongest relationship between racial resentment and attitudes against policies that could advantage blacks. As mentioned earlier, Kinder and Sanders (1996)

Although this scale does not stand out from the other three-item scales on other metrics, its superiority on the measures of predictive validity suggest its use. It correlates with the six-item scale at 0.92, has a coefficient α of 0.67 (close to the generally accepted 0.70 threshold), and test–retest correlation of 0.66 (the highest correlation of any subscale).[18] Taken together, we recommend against use of a two-item scale. Instead, we advocate using the three-item scale comprised of *deserve, slavery,* and *work up*.

Conclusion

Public opinion scholars face constraints on survey time and respondent patience. Scholars who seek to reduce the length of a scale must balance multiple (sometimes competing) criteria to identify the subscale that best meets their needs. Simple reliance on one statistic may not be sufficient to identify the optimally reduced scale. Here, we have provided a template that researchers can use to identify well-performing and efficient subscales of longer instruments, using only data from previous administrations of the full scale. Given that data from peer-reviewed publications are commonly available, researchers who seek to build upon published work (while implementing reduced versions of longer scales) can use our R package (ScaleReduce) and such readily available data to identify reliable, valid, and efficient reduced scales.

Our first empirical example found a four-item scale to be an appropriate alternative to the seven-item Risk Attitudes Scale. The revised risk scale is more parsimonious yet performs similarly to the seven-item scale in tests of both reliability and validity. In the second empirical example, we condensed the six-item Racial Resentment Scale into a three-item subscale, which performs similarly to the original scale.

More generally, we have proposed a framework for condensing any survey measure that has a multi-item scale, using comprehensive and systematic psychometric criteria of reliability and validity. Given the necessity of balancing efficiency and accuracy in designing survey questions to tap underlying phenomena that are challenging to measure, we hope our approach provides researchers with executable solutions to tackle this problem.

REFERENCES

Achen, Christopher. 1975. "Mass Political Attitudes and the Survey Response." *American Political Science Review* 69(4):1218–31.

Aronson, Elliot, Phoebe C. Ellsworth, J. Merrill Carlsmith, and Marti Hope Gonzales. 1990. *Methods of Research in Social Psychology*. Columbus, OH: McGraw–Hill Publishing Company.

Berinsky, Adam J., and Jeffrey B. Lewis. 2007. "An Estimate of Risk Aversion in the US Electorate." *Quarterly Journal of Political Science* 2(2):139–54.

Bizer, George Y., Jon A. Krosnick, Allyson L. Holbrook, S. Christian Wheeler, Derek D. Rucker, and Richard E. Petty. 2004. "The Impact of Personality on Cognitive, Behavioral, and Affective Political Processes: The Effects of Need to Evaluate." *Journal of Personality* 72(5):995–1028.

Blais, Ann-René, and Elke U. Weber. 2006. "A Domain-Specific Risk-Taking (DOSPERT) Scale for Adult Populations." *Judgment and Decision Making* 1(1):33–47.

noted that the two items dropped from the six-item scale explicitly discussed governmental programs, which could explain why the full scale has a stronger relationship with *other* governmental policies.

[18]In comparison, the two-item CCES scale correlates with the six-item scale at 0.86, has an α statistic of 0.46 and test–retest correlation of 0.61.

Bromiley, Philip, and Shawn P. Curley. 1992. "Individual Differences in Risk-Taking." Pp. 87–132 in J. Frank Yates, ed., *Risk-Taking Behavior*. Oxford, UK: John Wiley & Sons.

Byrnes, James P., David C. Miller, and William D. Schafer. 1999. "Gender Differences in Risk Taking: A Meta-Analysis." *Psychological Bulletin* 125(3):367–83.

Campbell, Donald T., and Donald Fiske. 1959. "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix." *Psychological Bulletin* 56(2):81–105.

Conover, Pamela J., and Stanley Feldman. 1986. "Emotional Reactions to the Economy: I'm Mad as Hell and I'm Not Going to Take it Anymore." *American Journal of Political Science* 30(1):50–78.

Delli Carpini, Michael X., and Scott Keeter. 1993. "Measuring Political Knowledge: Putting First Things First." *American Journal of Political Science* 37(4):1179–1206.

DeVellis, Robert F. 1991. *Scale Development: Theory and Applications*. Newbury Park, CA: Sage.

Dovidio, John F., and Samuel L. Gaertner, eds. 1986. *Prejudice, Discrimination, and Racism* (pp. 91–125). San Diego, CA: Academic Press.

Eckles, David L., Cindy D. Kam, Cherie L. Maestas, and Brian F. Schaffner. 2013. "Risk Attitudes and the Incumbency Advantage." *Political Behavior* 36(4):1–19.

Ehrlich, Sean, and Cherie Maestas. 2010. "Risk Orientation, Risk Exposure, and Policy Opinions: The Case of Free Trade." *Political Psychology* 31(5):657–84.

Fabrigar Leandre, Duan T. Wegener, Robert C. MacCallum, and Erin J. Strahan. 1999. "Evaluating the Use of Exploratory Factor Analysis in Psychological Research." *Psychological Methods* 4(3):272–99.

Feldman, Stanley. 1988. "Structure and Consistency in Public Opinion: The Role of Core Beliefs and Values." *American Journal of Political Science* 32(2):416–40.

Feldman, Stanley, and Leonie Huddy. 2005. "Racial Resentment and White Opposition to Race-Conscious Programs: Principles or Prejudice?" *American Journal of Political Science* 49(1):168–83.

Fischhoff, Baruch. 1992. "Risk Taking: A Developmental Perspective." Pp. 133–62 in J. Frank Yates, ed., *Risk-Taking Behavior*. Oxford, UK: John Wiley & Sons.

Gomez, Brad T., and J. Matthew Wilson . 2006. "Rethinking Symbolic Racism: Evidence of Attribution Bias." *Journal of Politics*, 68(3):611–25.

Hayton, James C., David G. Allen, and Vida Scarpello. 2004. "Factor Retention Decisions in Exploratory Factor Analysis: A Tutorial on Parallel Analysis." *Organizational Research Methods* 7(2):191–205.

Horn, John L. 1965. "A Rationale and Test for the Number of Factors in Factor Analysis." *Psychometrika* 32(2):179–85.

Hoyle, Rick H., Michael T. Stephenson, Philip Palmgreen, Elizabeth Pugzles Lorch, and R. Lewis Donohew. 2002. "Reliability and Validity of a Brief Measure of Sensation Seeking." *Personality and Individual Differences* 32(3):401–14.

Jacoby, William G. 1998. "Report on Values and Predispositions Items for the 1998 National Election Study." National Election Studies Technical Report Series. ftp 7.

Kam, Cindy D. 2012. "Risk Attitudes and Political Participation." *American Journal of Political Science* 56(2):817–36.

Kam, Cindy D., and Elizabeth Simas. 2010. "Risk Orientations and Policy Frames." *Journal of Politics* 72(2):381–96.

———. 2012. "Risk Attitudes, Candidate Characteristics, and Vote Choice." *Public Opinion Quarterly* 76(4):747–60.

Kinder, Donald R., and Allison Dale-Riddle. 2012. *The End of Race? Obama, 2008, and Racial Politics in America*. New Haven, CT: Yale University Press.

Kinder, Donald R., and Lynn M. Sanders. 1990. "*Mimicking Political Debate with Survey Questions: The Case of White Opinion on Affirmative Action for Blacks.*" Social Cognition 8(1):73–103.

Kinder, Donald R., and Tali Mendelberg. 2000. "Individualism Reconsidered: Principles and Prejudice in Contemporary American Opinion." Pp. 44-74 in David O. Sears, James Sidanius, and Lawrence Bobo, eds., *Racialized Politics: The Debate About Racism in America*. Chicago, IL: University of Chicago Press.

———. 1996. *Divided by Color: Racial Politics and Democratic Ideals*. Chicago, IL: University of Chicago Press.

Kinder, Donald R., and David O. Sears. 1981. "Prejudice and Politics: Symbolic Racism Versus Racial Threats to the Good Life." *Journal of Personality and Social Psychology* 40(3):414–31.

Kowert, Paul A., and Margaret G. Hermann. 1997. "Who Takes Risks? Daring and Caution in Foreign Policy Making." *Journal of Conflict Resolution* 41(5):611–37.

Lord, F. M., M. R. Novick, and A. Birnbaum 1968. *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

McConahay, John B. 1986. "Modern Racism, Ambivalence, and the Modern Racism Scale." Pp. 91–125 in John F. Dovidio and Samule L. Gaertner, eds., *Prejudice, Discrimination, and Racism*. San Diego, CA: Academic Press.

Montgomery, Jacob, and Josh Cutler. 2013. "Computerized Adaptive Testing for Public Opinion Surveys." *Political Analysis* 21:172–92.

Morgenstern, Scott, and Elizabeth Zechmeister. 2001. "Better the Devil You Know than the Saint You Don't? Risk Propensity and Vote Choice in Mexico." *Journal of Politics* 63(1):93–119.

Nadeau, Richard, Pierre Martin, and André Blais. 1999. "Attitude Towards Risk-Taking and Individual Choice in the Quebec Referendum on Sovereignty." *British Journal of Political Science* 29(3):523–39.

Peterson, Steven A., and Robert Lawson. 1989. "Risky Business: Prospect Theory and Politics." *Political Psychology* 10(2):325–39.

Podsakoff, Philip M., Scott B. MacKenzie, Jeong-Yeon Lee, and Nathan P. Podsakoff. 2003. "Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies." *Journal of Applied Psychology* 88(5):879–903.

Price, Vincent, and John Zaller. 1993. "Who Gets the News? Alternative Measures of News Reception and Their Implications for Research." *Public Opinion Quarterly* 57(2):133–64.

Romano, Jeanine L., Jeffrey D. Kromrey, Corina M. Owens, and Heather M. Scott. 2011. "Confidence Interval Methods for Coefficient Alpha on the Basis of Discrete, Ordinal Response Items: Which One, if Any, Is the Best?" *Journal of Experimental Education* 79(4):382–403.

Saris, Willem E., Melanie Revilla, Jon A. Krosnick, and Eric M. Shaeffer. 2010. "Comparing Questions with Agree/Disagree Response Options to Questions with Item-Specific Response Options." *Survey Research Methods* 4(1):61–79.

Schmitt, Neal. 1996. "Uses and Abuses of Coefficient Alpha." *Psychological Assessment* 8(4):350–53.

Smith, Gregory T., Denis M. McCarthy, and Kristen G. Anderson. 2002. "On the Sins of Short-Form Development." *Psychological Assessment* 12(1):102–11.

Sniderman, Paul M., Edward G. Carmines, Geoffrey C. Layman, and Michael Carter. 1996. "Beyond Race: Social Justice as a Race Neutral Ideal." *American Journal of Political Science* 40(1):33–55.

Stanton, Jeffrey M., Evan F. Sinar, William K. Balzer, and Patricia C. Smith. 2002. "Issues and Strategies for Reducing the Length of Self-Report Scales." *Personnel Psychology* 55(1):167–94.

Tarman, Christopher, and David O. Sears. 2005. "The Conceptualization and Measurement of Symbolic Racism." *Journal of Politics*. 77(3):731–61.

Tomz, Michael, and Robert Van Houweling. 2009. "The Electoral Implications of Candidate Ambiguity." *American Political Science Review* 103(1):83–98.

Tversky, Amos, and Daniel Kahneman. 1981. "The Framing of Decisions and the Psychology of Choice." *Science* 211(4481):453–58.

Vuong, Quang H. 1989. "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses." *Econometrica* 57(2):307–33.

Watson, David, and Lee Anna Clark. 1991. "The Mood and Anxiety Symptom Questionnaire." Unpublished Manuscript, Southern Methodist University, Dallas, TX.

Weber, Elke U., Ann-Renee Blais, and Nancy E. Betz. 2002. "A Domain-Specific Risk-Attitude Scale: Measuring Risk Perceptions and Risk Behaviors." *Journal of Behavioral Decision Making* 15(4):263–90.

Wiley, D. E., and J. A. Wiley. 1971. "The Estimation of Measurement Error in Panel Data." Pp. 364–74 in H. M. Blalock, ed., *Causal Models in the Social Sciences*. Chicago, IL: Aldine-Atherton.

Zaller, John. 1990. "Political Awareness, Elite Opinion Leadership, and the Mass Survey Response." *Social Cognition* 57(8):125–53.

Zeller, R. A., and E. G. Carmines. 1979. *Reliability and Validity Assessment*. Beverly Hills, CA: Sage.

Zuckerman, M. 1994. *Behavioral Expressions and Biosocial Bases of Sensation Seeking*. Cambridge, UK: Cambridge University Press.

Zuckerman, Marvin, Sybil Eysenck, and Hans J. Eysenck. 1978. "Sensation Seeking in England and America." *Journal of Consulting and Clinical Psychology* 46(1):139–49.

Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's website:

Table A1: Description of Data Sets
Table A2: Summary Statistics of Subscales, by Number of Items, Lab1
Table A3: Summary Statistics of Subscales, by Number of Items, KN2009
Table A4: Summary Statistics of Subscales, by Number of Items, ANES
Table A5: Summary Statistics of Subscales, by Number of Items, Lab2
Table A6: Summary Statistics of Subscales, by Number of Items, Lab3
Table A7: Summary Statistics of Subscales, by Number of Items, SSI2011
Table A8: Summary Statistics of Subscales, by Number of Items, YG2011
Table A9: Summary Statistics of Subscales, by Question, Lab1
Table A10: Summary Statistics of Subscales, by Question, KN2009
Table A11: Summary Statistics of Subscales, by Question, ANES
Table A12: Summary Statistics of Subscales, by Question, Lab2
Table A13: Summary Statistics of Subscales, by Question, Lab3
Table A14: Summary Statistics of Subscales, by Question, SSI2011
Table A15: Summary Statistics of Subscales, by Question, YG2011
Table A16: Risk Attitudes and Political Participation
Table A17: Policy Preferences and Probabilistic Outcomes
Figure A1: Correlations Between Risk Attitude Subscales and Known Covariates
Figure A2: Risk Attitudes and Future Participation
Figure A3: Risk Attitudes and Past Participation
Figure A4: Risk Attitudes and Preference for Probabilistic Policies
Figure A5: Subscale Correlations with Seven-Item Scale
Figure A6: Cronbach's α of Subscales
Figure A7: Test-Retest Correlations of All Subscales
Figure A8: Correct Classification Using Seven-Item Scale Benchmark
Table B1: Racial Resentment and Policy Preferences
Table B2: Summary Statistics of Subscales, by Number of Items, 1988 ANES
Table B3: Summary Statistics of Subscales, by Number of Items, 1990 ANES
Table B4: Summary Statistics of Subscales, by Number of Items, 1992 ANES