Political Analysis (2014) pp. 1–20
doi:10.1093/pan/mpt050

Extracting Wisdom from Experts and Small Crowds: Strategies for Improving Informant-based Measures of Political Concepts

Cherie D. Maestas
Department of Political Science, Florida State University, 531 Bellamy, Tallahassee, FL 32306
e-mail: cmaestas@fsu.edu (corresponding author)

Matthew K. Buttice
California Research Bureau, California State Library, Sacramento, CA 94237-0001
e-mail: matthew.buttice@library.ca.gov

Walter J. Stone
Department of Political Science, University of California, Davis, One Shields Ave., Davis, CA 95616
e-mail: wstone@ucdavis.edu

Edited by Jonathan Katz

Social scientists have increasingly turned to expert judgments to generate data for difficult-to-measure concepts, but getting access to and response from highly expert informants can be costly and challenging. We examine how informant selection and post-survey response aggregation influence the validity and reliability of measures built from informant observations. We draw upon three surveys with parallel survey questions of candidate characteristics to examine the trade-off between expanding the size of the local informant pool and the pool's level of expertise. We find that a "wisdom-of-crowds" effect trumps the benefits associated with the expertise of individual informants when the size of the rater pool is modestly increased. We demonstrate that the benefits of expertise are best realized by prescreening potential informants for expertise rather than post-survey weighting by expertise.

Many theoretical concepts of interest in political science lack direct empirical measures that are easily obtainable or comparable across units, leaving scholars to rely on conceptually distant or error-laden proxies in their analyses. One strategy to address this problem is the use of "expert" observers to rate targets of interest such as political actors, groups, contexts, or institutions. In political science, expert informants have been surveyed to measure bureaucratic agency placement on the left-right scale (Clinton and Lewis 2008), political party ideology in western democracies, (Castles and Mair 1984; Huber and Inglehart 1995; Benoit and Laver 2006; Hooghe et al. 2010; Albright and Mair 2011), party strategy and organizational characteristics (Kitschelt and Kselman 2012), common-space ideological positions of political actors (Saiegh 2009), and candidate qualities, prospects, and ideological positions in US House elections (Stone and Simas 2010; Stone et al. 2010; Buttice and Stone 2012). Seeking informed opinions is an attractive measurement strategy because researchers can tap the private, context-specific information held by knowledgeable individuals to create valid and reliable measures that are comparable across targets of interest, but this method is not without its challenges. The purpose of this article is to assess the consequences of fundamental design choices in these studies for the validity and reliability of measures created from aggregated informant judgments.

Authors' note: We would like to thank Lonna Rae Atkeson, Alex Adams, Ben Highton, Brad Jones, Chris Reenock, and the anonymous reviewers for helpful comments on previous drafts. Matthew Buttice began work on this project while at UC Davis and finished while at the California Research Bureau. The research results and conclusions expressed in this article do not necessarily reflect the views of the California Research Bureau or California State Library. Supplementary materials for this article are available on the *Political Analysis* Web site.

We use the terms "raters" and "informants" interchangeably to describe the individuals who rate a defined target of interest (e.g., a candidate, party, institution). The measures of targets that result from combining the individual judgments of raters are referred to as "target-unit" measures. In building such measures, researchers must identify the proper set and number of informants to survey as well as choose how to combine individual judgments into a single measure. Both choices are consequential for the quality of the target-unit measure.

One concern is with the trade-off between the level of expertise and the size of the informant pool that we suspect is inherent in most of these designs. The greater the expertise of individual informants, the higher the quality of their individual observations is likely to be. However, as informant expertise increases, the number of informants available to include in the study declines either because the number of highly informed experts is limited, or they are difficult to survey, or both. A further complication is that expert-informants' opinions are prone to errors that arise from individual biases, heuristics, or incomplete information (Powell 1989; Budge 2000; Steenbergen and Marks 2007; Whitefield et al. 2007; Albright and Mair 2011). Errors are particularly problematic for surveys that rely on a single informant to rate a target of interest, but are also problematic for studies that rely on small pools of informants per target.[1]

A potential remedy is to increase the size of the pool of informants per target, but doing so likely requires including informants who have less comprehensive knowledge about the targets of interest. The literature on the "wisdom of crowds" suggests there is value in seeking a larger, though less expert, pool of informants (e.g., Surowiecki 2004; Andersson, Edmund, and Eckman 2005; Sjoberg 2009). Numerous studies encourage the use of multiple rather than single informants (Phillips 1981; Boyer and Verma 2000), but to our knowledge no one has assessed the benefits of diluting expertise to increase the number of informants per target.

A second concern centers on combining informants' judgments. Once researchers have selected informants and solicited their responses, they face myriad choices for aggregating the individual responses to create valid and reliable target-level measures. Some studies promote post-survey weighting strategies to reduce the influence of respondent errors on inter-rater reliability (e.g., VanBruggen, Lillien, and Kacker 2002; Wagner, Lau, and Lindeman 2010). However, these studies generally rely on small pools of raters and stop short of evaluating whether post-survey aggregation weighting yields appreciable gains in the reliability and validity of the aggregated measures (but see Javaras, Godlsmith, and Laird 2011). We compare the gains from pre-survey informant selection based on expertise to the gains from post-survey weighting to determine which yields the greater benefit in the validity and reliability of the target-unit measures. We conclude that moderately expanding the size of the rater pool more than offsets the reduction in expertise of individual raters, and that prescreening raters for expertise is more beneficial than applying post-survey analytical weights for expertise.

1 Data Sources: 2010 US House Election Studies

Throughout the article, we use data from two sources: the 2010 UC Davis Congressional Election Study (UCD-CES), based on a national sample of 155 US House districts originally selected in 2006; and the 2010 Cooperative Congressional Election Study.[2] The UCD-CES study includes 2009 baseline and 2010 campaign surveys of two distinct groups of informants residing in the sample

[1] For a review of problems with single informant studies, see Phillips (1981).

[2] Maestas, Cherie D., Buttice, Matthew K., Stone, Walter J., 2013, "Replication data for: Extracting Wisdom from Experts and Small Crowds: Strategies for Improving Informant-based Measures of Political Concepts," http://dx.doi.org/10.7910/DVN/23170 IQSS Dataverse Network [Distributor] V1 [Version]. After drawing a random sample of one hundred districts in the contiguous forty-eight states, the sample was supplemented with a set of districts identified by multiple sources as open or likely competitive (*Cook Report*, *Congressional Quarterly*, *National Journal*, and *Sabato Crystal Ball*). Full details of the methodology of the UCD-CES Study can be found at http://electionstudy.ucdavis.edu/. Information about the Common Content of the 2010 Cooperative Congressional Election Study is available at http://projects.iq.harvard.edu/cces/.

districts: the "Delegate" pool composed of political elites defined by their official status as 2008 national convention delegates and state legislators who responded to the 2006 study; and the "YouGov" pool composed of opt-in panelists in the YouGov/Polimetrix online pool of survey respondents who were screened for their status as resident in one of the study districts and their expertise about and attentiveness to local and congressional politics.[3] The campaign-wave Delegate survey involved recontacting the same individuals who responded to the 2009 baseline study; the YouGov campaign-wave involved newly identified respondents who were screened using the same items in the 2009 baseline screening questions.[4]

The validity of using political elites such as state legislators and convention delegates to report on characteristics of congressional districts and candidates has been established in the Candidate Emergence studies spanning 1998–2006 (see Stone, Maisel, and Maestas 2004; Stone et al. 2010).[5] However, these studies also revealed limitations, namely a paucity of responses in some districts and substantial variation in the number of responses across other districts. The 2010 UCD-CES study sought to remedy this limitation by expanding the pool of raters to include YouGov panelists.

Opt-in Internet panels are well suited to providing large pools of potential informants because of the abundant number of panelists for which survey organizations collect profile information. Survey organizations collect information about panelists' place of residence along with a socio-demographic profile. As a result, it is relatively inexpensive to conduct a screening survey to identify those with the most expertise, a process we discuss in the next section.[6]

Table 1 shows that the variation in the size of rater pools is considerably greater for the delegate surveys than it is for the YouGov surveys, which reflects the greater control afforded by accessing opt-in panelists with an online survey. The goal in the UCD-CES was to identify twenty-five YouGov informants in each district and, on average, this technique was successful. The level of political expertise was higher among the Delegate respondents than the YouGov respondents but there were a greater number of YouGov respondents per district. Comparing the Delegate and YouGov informant studies is one way we can assess the effects of expertise and the number of informants on measures of reliability and validity.

To enhance the power of these comparisons, we also draw upon the 2010 Cooperative Congressional Election Study (CCES) "Common Content" survey (Ansolabehere 2010). The Common Content survey includes 12,844 respondents in the House districts included in the UCD-CES study. We leverage the large number of respondents per district in the 2010 CCES to create multiple pools of informants while varying the level of expertise in the pool and the size of the pool. Depending on the analysis, then, the comparisons we make are across repeated draws from the Common Content survey and between the Delegate and YouGov informant surveys. A drawback of the Common Content survey is that there are very few items that effectively ask respondents to act as "informants." For the small number of identical items on the Common Content and UCD-CES studies, we compare the results of the larger samples from the CCES to smaller but more expert samples of informants from the UCD-CES study.

[3]The expertise screening battery employed in the YouGov informant study is located at http://electionstudy.ucdavis.edu/files/MaestasButticeStone_Wisdom_Appendices.pdf. Potential respondents passed the screen and were administered the informant survey if they met at least five of six criteria: self-reported as highly informed about state government and politics; aware of the name of their House incumbent; expressed confidence in ability to answer questions about the district; and reported consulting news sources about politics and public affairs every day.

[4]The purpose of the baseline survey was to establish contact with the Delegate sample, and in both the Delegate and YouGov surveys to collect data on incumbent re-election prospects and other aspects of the strategic environment of the district before the campaign cycle began. The survey also included items designed to assess the ability of informants to report accurately on political, demographic, and economic aspects of their district. For the survey instruments, see http://electionstudy.ucdavis.edu/.

[5]For additional information on the Candidate Emergence Study, see http://ces.iga.ucdavis.edu/.

[6]A second important advantage of using an Internet survey of opt-in panelists over mail surveys of experts is the ability to elicit responses within a narrow time window. All responses for the online YouGov surveys were collected in a two-week window at the end of October 2010. Surveying of the delegates had to begin October 1 to allow time for the mail contact and response before Election Day.

Table 1 Response patterns to the delegate and YouGov district expert surveys

	Mean N *per district*	*Standard deviation of district* N	*Range of district* N	N *of districts*	N *of districts with panel* N = *1*
Baseline surveys 2009					
Delegate	8.4	4.2	1–22	153	2
YouGov	26.9	1.7	18–33	155	0
Campaign-Wave 2010					
Delegate	4.7	2.6	1–14	153	11
YouGov	26.7	1.8	18–32	155	0

2 Designing Informant Studies: Choosing the Size and Expertise of the Rater Pool

At face value, designing an informant-based study with a restrictive definition of expertise is attractive because, ceteris paribus, informants with greater expertise should yield answers with less error than those lacking expertise. However, it is not clear that small pools of highly expert raters trump larger pools of less expert raters when aggregating to the target-unit measure. If there is a trade-off between expertise and the size of informant pools available for study, the question is how to think about the implications of trading down the average expertise of raters for increased numbers of raters.

Individual informants, no matter how expert, are subject to both systematic and random errors (Powell 1989; Steenbergen and Marks 2007; Whitefield et al. 2007). In general, we think of random error as unproblematic when aggregating observations of large groups of individuals because random errors cancel out in the aggregate. Likewise, individual systematic error may cancel out in the aggregate if individual biases are randomly distributed across respondents. However, when error in judging a target is not independent across informants, such as when all individuals assessing the target hold similar social identifications (e.g., party identification), or when informants directly collaborate when forming judgments, the expected value of aggregated target-unit error is no longer zero. Systematic errors that persist after aggregation can invalidate the overall target-unit measure if they alter the ordinal and/or cardinal ranking of targets relative to one another.

One strategy to improve the quality of informant-based measures is to increase the number of diverse and independent observers of a target. Indeed, the rationale for multi-rater studies rests on the premise that increasing the number of informants reduces the influence of individual errors on the aggregated measure (Boyer and Verma 2000). The "wisdom-of-crowds" logic (Surowiecki 2004) highlights how aggregating the independent judgments or predictions of a diverse set of individuals yields a high degree of accuracy, even when each individual has imperfect information. When the errors in informant ratings are independent, the variance of the target-unit measure (i.e., aggregated informant ratings) declines with the square root of N for any level of informant expertise.

Comparisons of public and expert forecasts of election outcomes, sporting events, and economic conditions demonstrate that it is possible for partially informed publics to know more, collectively, than single or small groups of experts (Surowiecki 2004; Andersson, Edmund, and Eckman 2005; Sjoberg 2009). This wisdom-of-crowds logic implies that increasing and diversifying the pool of informants for a given target will increase the validity and reliability of target-unit measures by reducing the influence of individual-level response errors on the aggregated quantity of interest. This perspective stands in contrast to the more typically used and intuitive approach of restricting the sampling frame to a small number of highly expert individuals.

Where possible, we suggest that it is also important to identify and correct for sources of systematic bias related to expertise. Political elites or other politically involved observers tend to hold biased views due to strong pre-existing political identifications and, therefore, are likely to provide biased judgments. Politically engaged party activists, for example, tend to rate the qualities and skills of incumbents that share their party affiliation more highly than they rate those same qualities for the opposition party (Stone et al. 2010). For studies that rely on one or a few elite raters, perceptual bias can distort the collective rating of an individual target. One alternative to reduce

this problem is to model and purge known sources of bias prior to aggregation (see Stone et al. 2010; Buttice and Stone 2012). In the next sections, we compare target-level measures built from data that are both purged and unpurged of known systematic biases for different-sized rater pools and different levels of expertise.

2.1 *Examining the Trade-Off between the Size and Expertise of Rater Pools*

To examine the effects of the composition and size of the rater pool on aggregated measures, we need responses to an informant-based measure that can be validated externally. The best option from the 2010 studies is the assessment of US House incumbents' ideology, rated on a seven-point liberal-conservative scale. We can compare the target-unit measures aggregated from our respondent pools to incumbents' ideological positions as measured by the well-validated DW-NOMINATE measure (Poole and Rosenthal 1997). We create two target-unit measures (i.e., House incumbent measures) by obtaining the district mean of each House incumbent's ideology, one based on the Delegate pool responses and one based on the YouGov pool responses. We also obtain the district-level variances for each in order to calculate a measure of reliability of the target-unit measure developed by Jones and Norrander (1996), a measure of reliability for aggregated survey data:

$$\hat{\mathrm{Ep}}^2 = \frac{\sigma^2(\mathrm{a})}{\sigma^2(\mathrm{a}) + \frac{\sigma^2(\mathrm{j:a,\,e})}{\mathrm{n_j}}}. \tag{1}$$

In this formula, n_j is the number of respondents in district j, $\sigma^2(a)$ is the variance of the mean across all districts, and $\sigma^2(j:a,e)$ is the variance of individual responses within districts. This measure takes into account the degree of variation both within and between districts; greater reliability occurs when the ratio of between-unit variation to within-unit variation is high. In other words, when variation in raters' judgments of a common target is low, and the discrimination between targets is high, the target-unit measure is more reliable. In addition to assessing reliability, we assess the validity of the target-unit measure of ideology by correlating it with first-dimension DW-NOMINATE scores for incumbents in the study.

The reliabilities of the informant-based incumbent ideological placement variable in the Delegate versus YouGov studies are 0.93 and 0.99, respectively. When we compare the correlations between the informant-based measure of incumbent ideology and DW-NOMINATE scores in the two studies, the high correlation in the Delegate study ($r = 0.92$) is even stronger in the YouGov survey ($r = 0.96$).[7] These reliabilities and correlations are high by any standard, and suggest the utility of informant-based measures. However, the YouGov survey is based on larger district pools of less expert informants than is true of the Delegate survey. Despite the fact that individual YouGov respondents score lower on expertise measures than the Delegate respondents, the reliability and validity of the YouGov measures are greater than the measures based on Delegates. The question is whether there is general evidence supporting the use of larger, less expert, pools of informants.

To investigate the trade-off between the size of rater pools and the expertise of individual informants in a more systematic way, we turn to a different mode of analysis designed to impose greater control over our analysis. The CCES Common-Content survey includes questions asking respondents to place the Democratic and Republican candidates in their district on a seven-point

[7]The correlations between DW-NOMINATE scores and informant-based measures are not due simply to party polarization but they are stronger when pooling Democrats and Republicans in the analysis because the within-party variance on all candidate-position measures is greatly reduced, compared with the full sample. However, the key comparison from which we derive support for our argument—the difference in correlations between the YouGov respondent pool and the Delegate pool—remains unchanged whether we analyze the data within or across parties. For the within-party analysis for Republicans, we drop the case of Ron Paul because he was an extreme outlier in DW-Nominate scores. We also drop those cases in which we had fewer than two raters per target. The within-party comparisons of correlations show that target measures built from smaller pools of Delegate raters do no better than measures built from larger pools of less expert YouGov raters: Democratic incumbents ($N = 90$)—Delegates, $r = 0.71$, $N = 90$; YouGov, $r = 0.71$; Republican incumbents ($N = 41$)—Delegates $r = 0.40$, YouGov, $r = 0.53$.

liberal-conservative scale. This is a standard perceptual item, which can be taken as asking ordinary mass-survey respondents to act as "informants" about the ideological positions of the candidates running in their district.[8] These perceptual items can be aggregated by taking the mean rating within districts as estimates of candidates' positions in exactly the same way that we aggregate any informant rating of a candidate characteristic. Because the district samples in the Common Content study were large and the level of political expertise of respondents varied within each district, we can use them to simulate estimates of candidates' positions based on informant pools of varying size and expertise.

The districts for the simulations were selected to overlap with districts in the UCD-CES study to facilitate comparisons with results from the Delegate and YouGov surveys.[9] We focus on examining the effects of size and expertise of relatively small pools of informants ranging from $N = 2$ to $N = 30$. We leverage the naturally occurring variation in awareness and expertise among CCES respondents to construct pools of raters randomly selected from all CCES respondents in the districts, compared with respondents who were selected for their political expertise. We assume that those with the greatest interest and knowledge were also more expert when placing their House incumbent on the liberal-conservative scale. The expertise selection method used respondents' general political knowledge, interests in politics, news consumption, political participation, and an indicator of whether they accurately recalled the party of their US House incumbent (scale reliability of 0.78). Each item is standardized and the index is scored on a 1–100 scale. The mean expertise score for all respondents is 72 (SD = 15); for those who passed the expertise screen the mean expertise score is 83 (SD = 6). Descriptions of the expertise questions are included in the supplementary appendix. Individuals passed the expertise screen if they scored in the top 50th percentile of this expertise index. We explored a range of cutoffs for expertise and found no major differences in our conclusions.

We drew two subsets of raters, with replacement, from each district's population of respondents, varying the size of the rater pool from 2 to 30: one set was from all respondents in the district and one was drawn only from respondents in each district who passed the expertise screen. Each draw for each condition defined by N and expertise was repeated one thousand times. We required that districts have at least thirty raters who passed the expertise screen to avoid the possibility that the observed effects of increasing the pool N are confounded with the number of experts in the district sample. For each rater pool in each district, we calculated the mean and variance of incumbents' ideological placements.

Figure 1 presents the average reliability and average validity of incumbent ideological placements by rater pools for different-sized N. In general, substantial payoffs accompany even modest increases in the size of informant pools, whereas screening for expertise has the greatest effect on small rater pools. When the rater pool is small the effect of expertise improves both the average Jones-Norrander reliability of the incumbent ideological position measure (left panel) as well as increasing the validity—the average correlation between the DW-NOMINATE measure and the simulated informant-based measures (right panel). When the informant pool contains only two respondents per district, the average reliability of the incumbent ideology measure improves from 0.73 to 0.82 when raters are prescreened for their expertise. However, the gap in reliabilities narrows between the expertise-screened rater pool and the randomly drawn rater pool as the number of raters per target grows. Increasing the sample size by modest amounts (i.e., to five or ten respondents) reaps substantial gains, even when informants are drawn without regard to their expertise. For example, target-unit measures built from pools of twenty randomly drawn raters exceed the reliability of those built from pools of nine or fewer expertise-screened raters. Indeed, the average reliability for the target-unit measure exceeds 0.9 with as few as ten randomly selected raters per target.

[8]The wording of the Common Content candidate-placement items is identical to items asked on the Delegate and YouGov surveys.

[9]The districts are not a random selection of districts in the UCD-CES study because the total number of respondents as well as the number of respondents who met the expertise criterion varies across districts. All of the simulations are reported on the same subset of districts to control for possible district-level confounds.

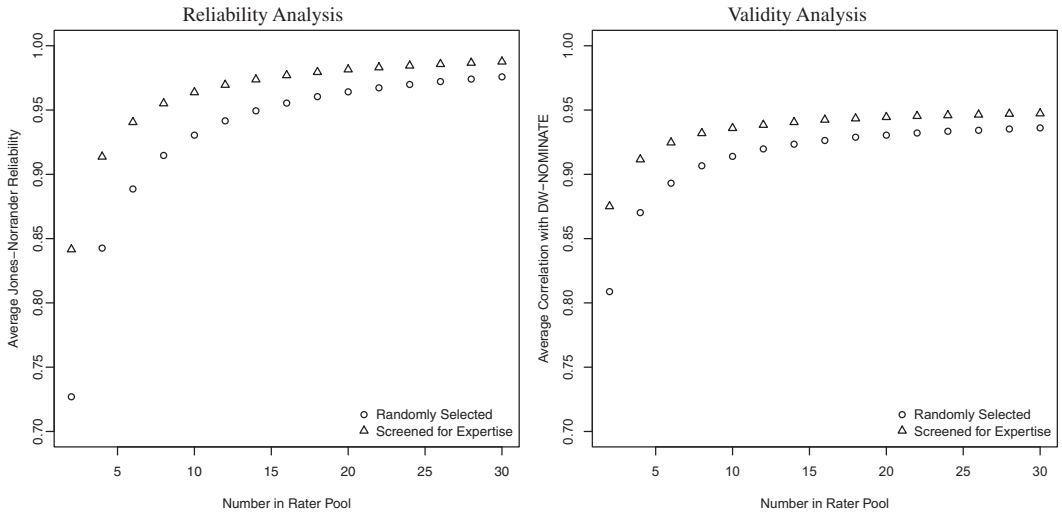

Fig. 1 Effects of size of rater pool and rater expertise on reliability and validity of aggregate incumbent position measures. Each point represents the mean reliability or correlation with DW-NOMINATE of aggregate measures based on one thousand samples, with replacement, from CCES respondents resident in a fixed set of districts (N of districts = 116).

Expertise screening also improves the correlation of the informant-based measure with the DW-NOMINATE score when the size of the rater pool is small (right panel). With two respondents per district, limiting the pool to expert informants increases the average correlation from 0.83 to 0.89. As with reliabilities, however, larger gains in validity come from expanding the rater pool rather than screening for expertise. Consider a comparison between measures based on five expert raters per district and twenty-five random raters per district. The average correlation of the measure with DW-NOMINATE based on five raters selected for their relatively high expertise is 0.92. However, the correlations from pools set at twenty-five randomly selected raters (i.e., without the expertise screen) produce a mean correlation of 0.95. In short, the gains in the quality of the target-unit measure from restricting the pool to "experts" swiftly declines as N increases. This supports the wisdom-of-crowds logic, which relies on combining multiple observations of individuals with imperfect information. A diverse group of informants of sufficient size can generate aggregate assessments close to or better than those produced by small groups composed of individuals with markedly greater expertise (Sjoberg 2009).

Of course, it is possible that these effects are context-dependent. The preceding analysis highlights average results across all incumbent and district contexts. Does the same result hold for incumbents who are difficult to rate? We reran the simulations, splitting the districts/incumbents into two groups based on how far the incumbent's DW-NOMINATE score was from the median DW-NOMINATE score of his or her party. We expect incumbents who are further from their party's median to be more difficult to rate because informants infer the incumbent's position based on the party position (Feldman and Conover 1983; Dancy and Sheagley 2013). The bottom 50th percentile on the distance variable formed the "close to the party median" group and the top 50th percentile formed the "far from the party median" group. We ran one thousand draws and estimated the correlation between the simulated informant-based measure of ideology and DW-NOMINATE for the "close to the party median" group and the "far from the party median" group separately. We also estimated the Jones-Norrander reliability for each group.

Figure 2 shows patterns quite similar to those in Fig. 1 where the target-unit measures from screened experts outperform the target-unit measures from randomly selected respondents when the N is small regardless of whether the incumbent is easy or difficult to rate. As in the earlier simulations, target measures built from larger pools of raters—expert or random draws—perform better than measures that rely on only a handful of expert raters. However, measures from both expertise-screened and random draws of raters perform more poorly for incumbents who take ideological

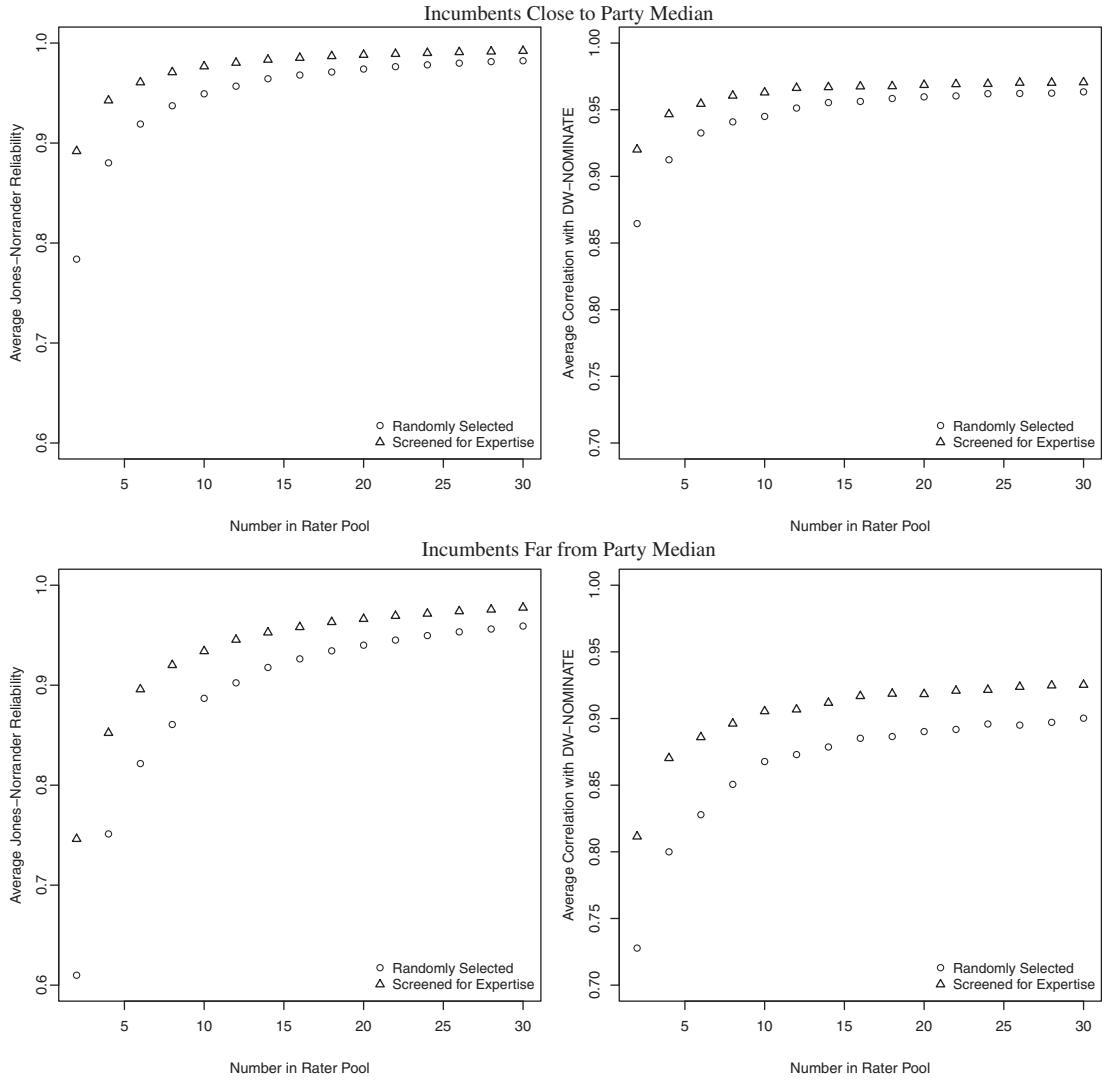

Fig. 2 Effects of size of rater pool and rater expertise on reliability and validity of aggregate incumbent position measures as incumbents' distance from party median changes. Each point represents the mean reliability or correlation with DW-NOMINATE of aggregate measures based on one thousand samples, with replacement, from CCES respondents resident in a fixed set of districts (N of districts = 116).

positions far from the party median. The correlations between the target-unit measures and DW-NOMINATE are lower across the board in the bottom right panel compared to the top right panel, as are the reliabilities in the lower left panel. The analysis suggests that experts have an advantage in these difficult-to-rate situations. Even so, a pool of about twenty randomly selected raters performs better than small pools of screened raters (i.e., $N < 7$). Ideally, in hard-to-rate situations, expanding the pool of expert raters may be the best choice, but in the absence of available experts, augmenting the pool with lower-expertise informants yields appreciable gains in the quality of the target-unit measures.

2.2 *Comparing Expertise and Target-Unit Measures from the 2010 Election Studies*

To facilitate validation checks of informants' expertise for the three different informant studies, we compare questions that have either objectively correct answers or strong external criterion validity variables. Table 2 presents the average individual-level error of the delegate and YouGov rater-pools for five items. Since three of these questions were also asked on the CCES Common Content,

Table 2 Mean square error of individual judgments, by rater pool

	Delegate	*YouGov*	*CCES*
Incumbent ideology	1.62	1.90	2.91
Senator 1 ideology	3.38	3.95	5.48
Senator 2 ideology	2.56	3.15	4.25
2010 Winning House Candidate	2.87	3.22	
2008 Presidential Vote Share	2.33	3.00	

Note. Cell entries represent the mean square error of responses for each rater pool. The relevant external criterion variables were regressed on individual-level responses separately for each group, and the residuals from the regression were squared.

Table 3 Reliability and validity of aggregate measures, by rater pool

	Delegate pool		*YouGov pool*		*CCES*	
	Average N = 4.7		*Average N = 26.7*		*Average N = 126.6*	
	Reliability	*Correlation*	*Reliability*	*Correlation*	*Reliability*	*Correlation*
Incumbent ideology	0.93	0.92	0.99	0.96	0.99	0.95
Senator 1 ideology	0.92	0.88	0.98	0.95	0.99	0.95
Senator 2 ideology	0.92	0.89	0.98	0.97	0.99	0.97
2010 Winning Candidate	0.73	0.77	0.94	0.91		
2008 Presidential Vote Share	0.74	0.79	0.94	0.93		

Note. Cell entries represent the aggregate reliability and correlation with external criterion variables for each item, by rater group. The average *N* reported for each rater pool is the district-level average of potential observations. There is variation in the actual *N* across items due to missing data.

we include the average error of all CCES respondents in the same districts. The data indicate exactly what we would expect: the Delegate study was based on individual informants with the greatest expertise (lowest mean square error), followed by the YouGov informants, with the Common Content respondents least expert.[10]

Table 3 reports the reliability and validity of the target-unit measures based on the district-level means from each of the three informant studies. The results show that despite the greater individual-level accuracy of the delegate respondents, the *target-unit* level measures from both the YouGov and CCES outperform measures based on the Delegate survey. This result is consistent with the simulation results in Fig. 1. The average size of the rater pool was substantially larger for the latter two, leading to stronger aggregate measures compared to the smaller yet more expert delegate pool. It is worth noting that despite the much larger *N* in the Common Content (nearly five times the average size of the YouGov pool), the reliabilities and correlations with the criterion variable were essentially identical. This suggests that the same quality of measures offered by a large number of district respondents can be achieved by a smaller pool of prescreening opt-in panelists, and that a small panel of prescreened opt-in panelists can equal or improve upon measures based on hard-to-survey local political elites.[11]

In Table 4, we also examined whether districts or incumbent characteristics influenced the degree of error in the target-unit measures by using district and incumbent characteristics to predict the squared error from regressing the YouGov and Delegate target-unit measures on the criterion DW-NOMINATE scores. The model is based on Delegate and YouGov measures in each district to test

[10]All pairwise comparisons were significant at the 0.10 level; all but two were significant at the 0.05 level.

[11]Given how well the Common Content measures perform, we might ask why bother with informant studies at all? One reason is that informant surveys often include a great many items that tap relatively specialized knowledge. It would be expensive, impractical, or impossible to include many of these items on a general mass survey, and in some cases, the information would be so lacking among mass respondents that the wisdom-of-crowds logic would not apply (Page 2007). In the case of the 2010 study, it would have been impractical to include the 150 items on the baseline and campaign informant studies on the Common Content study.

Table 4 Explaining error in target-unit ideology measures with incumbent, district, and rater pool characteristics

	Model 1		*Model 2*		*Model 3*	
Absolute DW-NOMINATE score	−1.10**	(0.52)	−0.98*	(0.52)	−0.50	(0.68)
Absolute distance from party median	2.72**	(0.59)	2.83**	(0.58)	2.35**	(0.76)
Standard error of DW-NOMINATE	4.37**	(1.60)	4.15**	(1.58)	4.61**	(2.07)
Challenger spending (logged)	−0.07	(0.04)	−0.05	(0.04)	−0.01	(0.05)
District competiveness	−0.50	(0.48)	−0.54	(0.47)	−0.02	(0.62)
Incumbent's terms in office	0.02	(0.01)	0.01	(0.01)	0.01	(0.02)
Delegate pool	0.35**	(0.11)	−1.47**	(0.57)	−1.01*	(0.60)
Number of informants			−0.08**	(0.03)	−0.01	(0.04)
Delegate pool × absolute DW-NOMINATE score					−0.91	(0.89)
Delegate pool × absolute distance from party median					0.85	(1.00)
Delegate pool × standard error of DW-NOMINATE					−1.22	(2.73)
Delegate pool × challenger spending (logged)					−0.07	(0.07)
Delegate pool × district competiveness					−1.04	(0.82)
Delegate pool × incumbent's terms in office					0.01	(0.02)
Delegate pool × number of informants					−0.11**	(0.05)
Constant	0.28**	(0.08)	1.19**	(0.29)	0.36	(0.50)
Observations	262		262		262	
Number of districts	131		131		131	

*$p < 0.10$, **$p < 0.05$

Note. Dependent variable is the squared residual from a regression of the informant-based incumbent ideology measure on DW-NOMINATE scores. All independent variables, with the exception of "Delegate Pool," were mean centered to make interpretation of the results in Model 3 more manageable. Multilevel model estimates with standard errors in parentheses.

whether district and incumbent characteristics have a differential effect on the degree of error for measures based on experts and screened raters. Therefore, we used a multilevel model that allowed the constant to vary across districts to control for the fact that we have two measures (one based on the Delegate pools, one based on the YouGov pools) for each district.[12]

We included in the regression the absolute DW-NOMINATE score to capture extremity, the absolute distance of incumbent from the party median, and the standard error of the DW-NOMINATE scores.[13] The latter two measures tap unpredictability, in the first instance because informants should have more difficulty judging the positions of incumbents who position themselves far from their party, and in the second instance because the larger the standard error of the NOMINATE score, the lower its precision. Models 1 and 2 show that all three measures influence errors in ways we might expect. Informants do a better job placing incumbents who are more extreme in their ideology and do worse in predicting those who are atypical of their party or inconsistent in their voting records. We also included the number of terms the incumbent has served in office and measures of district competitiveness and logged challenger spending since it seemed plausible that campaigns might raise the visibility or provide increased signals about incumbent ideology, but none of these covariates affected squared errors.[14] Finally, we investigated whether expert raters did better in circumstances where rating targets was difficult by including interactions between the source of the raters and each incumbent and district characteristic in Model 3, but found none to be significant.[15]

[12]We have run this analysis on target unit measures built from ratings unpurged of the partisan bias we describe later. As a robustness check, we ran an identical analysis using rater data purged of partisan bias to construct the target-unit measures and had substantively identical results.

[13]This measure is a bootstrapped standard error of the DW-NOMINATE score. See Carroll et al. (2009) for a full discussion of its calculation.

[14]We use the Huckfeldt et al. (2007) measure of district competitiveness, $4p(1-p)$, where p is the proportion of the two-party vote cast for the Democratic candidate (Huckfeldt et al. 2007).

[15]We examined the marginal effect of the delegate pool dummy variable across the range of each incumbent and district characteristic. With the average rater pool size found in the delegate panels ($N = 5$), the improvement associated with

It is worth noting that the coefficient for the "delegate pool" dummy variable was *positive* in Model 1, suggesting that delegate-based measures had greater error, on average, than YouGov measures. However, this is largely an artifact of the size of the rater pool, which in the case of delegates is smaller. Once size of the informant pool is controlled in Model 2, it is clear that expert raters do better on average. However, independent of expertise, an increase in the number of raters reduced error in the target-unit measures.

Taken together, the results imply that scholars can expand the size of the rater pool a moderate amount, even if the raters are slightly less expert, and achieve the same or greater reliability and validity of target-unit measures. Further, the gains from adding raters quickly levels off. Measures based on pools of twenty to thirty raters prescreened for their expertise perform as well or better than those based on more than one hundred mass-respondent raters. Surveying less expert but more accessible and larger pools of potential informants may be a more cost-effective strategy than approaching highly expert but limited pools of informants.

3 Evaluating Post-Survey Strategies for Improving Validity and Reliability

In this section, we evaluate post-survey strategies for reducing errors through purging systematic bias and weighting for expertise. We begin by examining the reliability and validity of incumbent ideology measured through simulation of informant pools from CCES respondents.

3.1 *Purging Systematic Bias*

In the study of congressional elections, individual-level systematic bias is a concern because politically aware informants tend to hold strong partisan identifications that influence their views of incumbents. Informants may also engage in wishful thinking when judging the prospects and qualities of an incumbent from their own party. Known systematic biases can be addressed in the design of the study through the intentional selection of informants with offsetting biases, which the UCD-CES survey attempted by contacting equal numbers of Republican and Democrat delegates. However, in small pools of informants or in pools where Democrats and Republicans respond in unequal numbers, the biases may not offset, so measures may be improved by estimating and purging bias from the individual-level measures prior to aggregation. In purging the data, we assumed that independents placed candidates without bias. We subtracted the average partisan bias from independent placements prior to aggregating responses:

$$\text{trait response}_{ij} = \beta_{0j} + \beta_{1j}(\text{pid}_{ij}) + \varepsilon_{ij}, \tag{2}$$

where

- i indexes informants and j indexes districts;
- $\beta_{0j} = \gamma_{00} + \delta_{0j}$;
- $\beta_{1j} = \gamma_{10} + \delta_{1j}$;
- γ_{00} and γ_{10} denote fixed district-level parameters;
- ε_{ij}, δ_{0j}, δ_{1j} are normally distributed error terms; and
- pid_{ij} is a three-point variable coded −1 for Democrats, 0 for independents, and 1 for Republicans.

The intercept, β_{0j}, varies across congressional districts, reflecting the fact that informants in different districts observe different candidates. Because the degree of partisan bias may vary across districts, we allow the slope, β_{1j}, to vary as well. The candidate characteristics for a given item are calculated by subtracting the average partisan bias in the district (β_{1j}) from informant responses and averaging across all informants in the district:

$$T_j = \frac{1}{n_j} * \sum_{i=1}^{n_j} (\text{trait response}_{ij} - \beta_{1j}(\text{pid}_{ij})). \tag{3}$$

the delegate-based measures, once the size of the informant pool is controlled, is not significantly greater for difficult-to-rate incumbents than it is for those who are easier to rate.

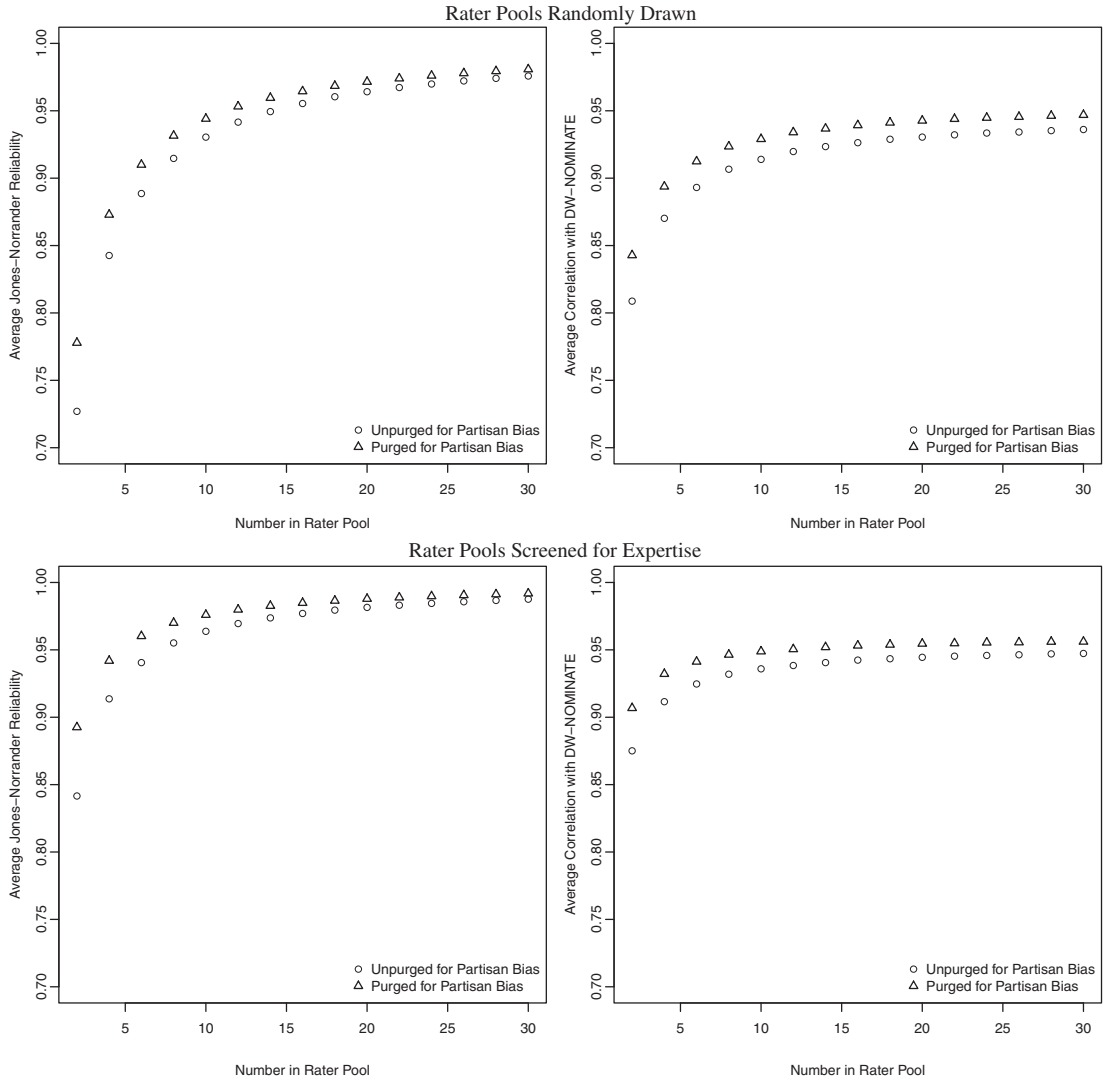

Fig. 3 Effects of bias-correction and rater expertise on reliability and validity of aggregate incumbent position measures. Each point represents the mean reliability or correlation with DW-NOMINATE of aggregate measures based on one thousand samples, with replacement, from CCES respondents resident in a fixed set of districts (N of districts = 116).

n_j equals the number of informants in district j and candidate trait (T) is the estimated value of candidates' issue positions, personal characteristics, or prospects of re-election in district j.

Figure 3 illustrates the effects of purging individual estimates of House incumbent ideology as the size of rater pools increases. The top two panels present reliabilities and validity correlations when informant pools are randomly selected from all Common Content respondents in the districts, whereas the bottom row presents results among CCES respondents who scored above the median on the expertise index in supplementary appendix 1.

It is apparent that the gains from purging are strongest for small informant pools. For modest-sized pools of informants, the Jones and Norrander reliability of the target-unit measure is nearly identical for both expert and random pools of raters (an average of 0.97 for the unscreened pools and 0.98 for those pools screened for expertise). However, an appreciable, albeit small, advantage remains for the validity of the target-unit measure for both screened and randomly selected rater pools. The average correlation between DW-NOMINATE and the informant-based ideology

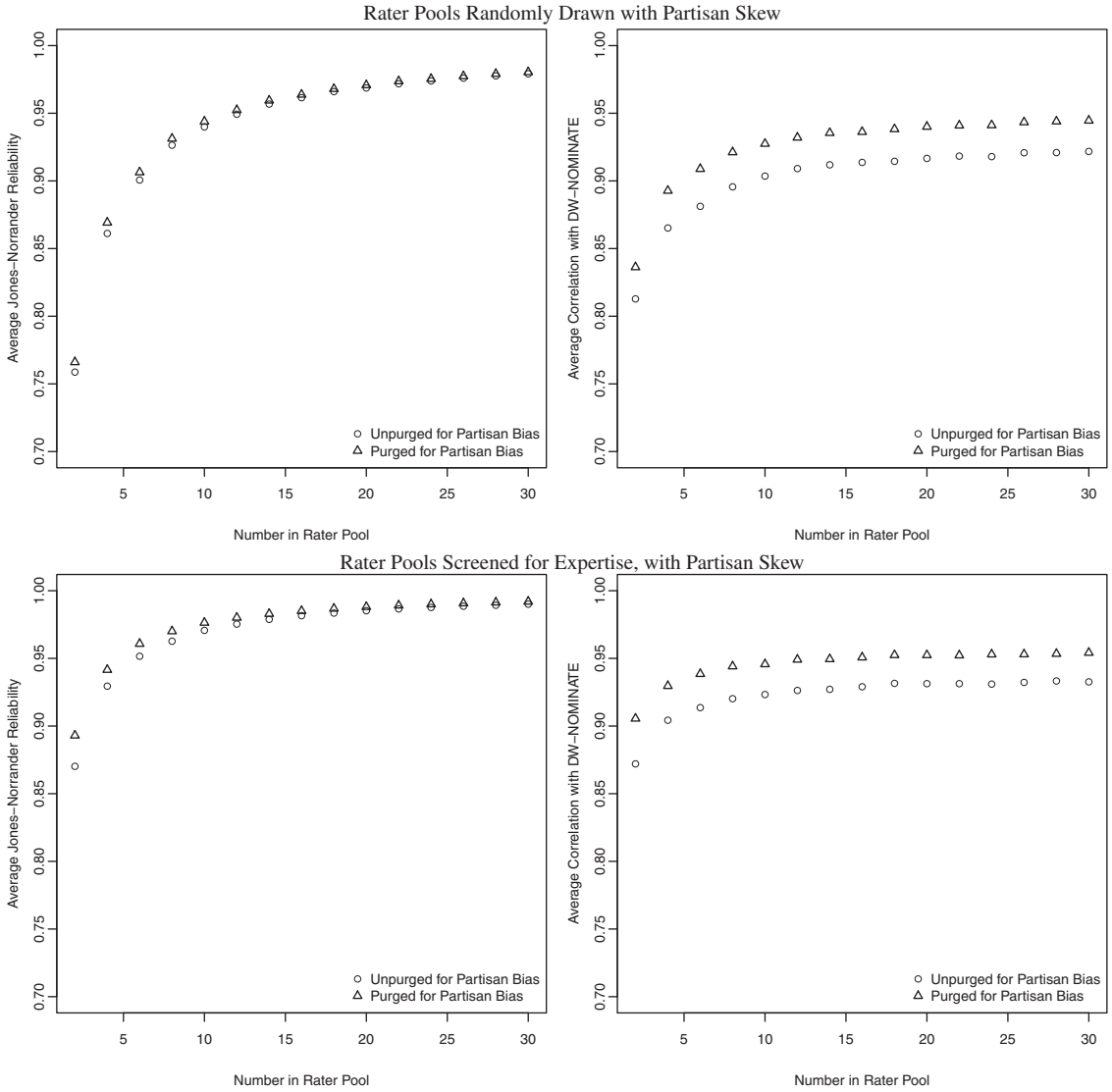

Fig. 4 Effects of bias-correction and rater expertise on reliability and validity of aggregate incumbent position measures with skewed partisan distribution of informants. Each point represents the mean reliability or correlation with DW-NOMINATE of aggregate measures based on one thousand samples, with replacement, from CCES respondents resident in a fixed set of districts (N of districts = 116).

measure is higher when informant opinions were purged of bias prior to aggregating for all sample sizes in the range we examine.[16]

The simulations run previously take a random draw of partisans and report averages. But, individual researchers may be unlucky in a single draw and obtain samples that skew heavily toward one party or the other. Is purging sufficient in such cases? We ran one thousand simulations where we forced a skewed partisan distribution in each draw in a random half of the districts to skew 75% Republican, 10% Independent, and 15% Democrat, and 15% Republican, 10% Independent, and 75% Democrat in the other half. Further, we ran separate analyses for randomly selected rater pools and rater pools prescreened for expertise. This permits us to see whether purging is more important to one type of rater pool than the other. The simulation results in Fig. 4 show that expert-based

[16]The small effect in Fig. 4 reflects the minimal bias in informant perceptions of incumbents' ideological positions. In Monte Carlo simulations not presented here, the effect of purging increases when applied to items with greater bias.

measures produce both reliable and valid target-unit measures, even when distributions of raters are heavily skewed. However, in such circumstances, preprocessing the data to purge bias prior to aggregating is an important step to increasing the quality of the measure. Although the skew matters slightly less for experts than for randomly drawn raters, the difference is not large. Experts achieve a higher correlation and reliability at smaller *N*s than the less expert rater pools but as *N* increases, both pools of raters produce quality measures.

3.2 *Weighting Individual Responses by Expertise*

A critique of using simple mean aggregation to combine informant observations is that it equally weights all individuals regardless of their level of expertise, whereas weighting by expertise during aggregation can increase the average inter-rater agreement per target (VanBruggen, Lilien, and Kacker 2002; Wagner, Rau, and Lindemann 2010). Proponents of this approach assume that increasing inter-rater agreement also serves to increase the validity and reliability of the target-unit measure. In contrast, the "wisdom of crowds" logic suggests that with a sufficient number of diverse informants simple unweighted aggregations of public judgments are highly accurate (Gaissmaier and Marewski 2011).

We believe that the conflicting advice from the two literatures reflects the difference in the size of pools examined. Studies that emphasize weighting by expertise tend to focus on small pools of raters, usually five or fewer. Studies emphasizing the wisdom-of-crowds effect rely on larger pools of informants and compare small expert pools to large public pools (e.g., Sjoberg 2009). What is missing from both is a comparison of the benefits of expertise-weighting when the number of respondents is controlled.

We ran simulations comparing the validity and reliability of target-unit measures built from expertise-weighted informants and unweighted informants for various sized informant pools. We weight each informant by the same items used in the CCES expertise screen described earlier and in the supplementary appendix. The weighting scheme draws upon approaches developed by VanBruggen, Lilien, and Kacker (2002) that have been found to be both computationally simple and similar in effectiveness to more complex strategies such as Bayesian estimation. The expertise-weighted mean of candidate traits (wT) is based on the following formula:

$$\mathrm{w}T_j = \sum_{j=1}^{n_i}\left[\frac{\mathrm{expertise}_{xij}^{\propto}}{\sum_{j=1}^{n_i}\mathrm{expertise}_{xij}^{\propto}}\ t_{ij}\right]. \tag{4}$$

In this case, *i* informants rate attribute *t* for target *j*. The rating of attribute *t* in the weighted mean is proportional to the relative expertise of rater *i* to the expertise of the pool rating *j*. This factor is scaled by α, a parameter that varies the strength of the expertise weight.

The top two panels in Fig. 5 report simulated informant pools using all CCES respondents in the districts, whereas the bottom two panels report simulations limiting the informant pools to those who passed the expertise screen. As before, the curves measure mean reliability of incumbent placements using Jones-Norrander reliability, and validity, the mean correlations of incumbent placements with DW-NOMINATE scores.

Figure 5 shows that both the reliability and validity of the informant-based measure improve when expertise weights are employed among mass-respondent raters not screened for their expertise. However, the bottom two panels demonstrate that prescreening Common Content respondents for their political expertise eliminates the benefit from post-survey weighting. The greatest gains in reliability and validity stem from increasing the size of the informant pool, although there is a modest payoff associated with expertise weights among raters not preselected for their expertise.

4 Substantive Example from UCD-CES Informant Studies

Thus far, we have focused our analysis on measures of incumbent ideology because we have a clear criterion variable in DW-NOMINATE scores. However, the purpose of developing informant-based studies is to tap measures that are not readily available through alternative sources and,

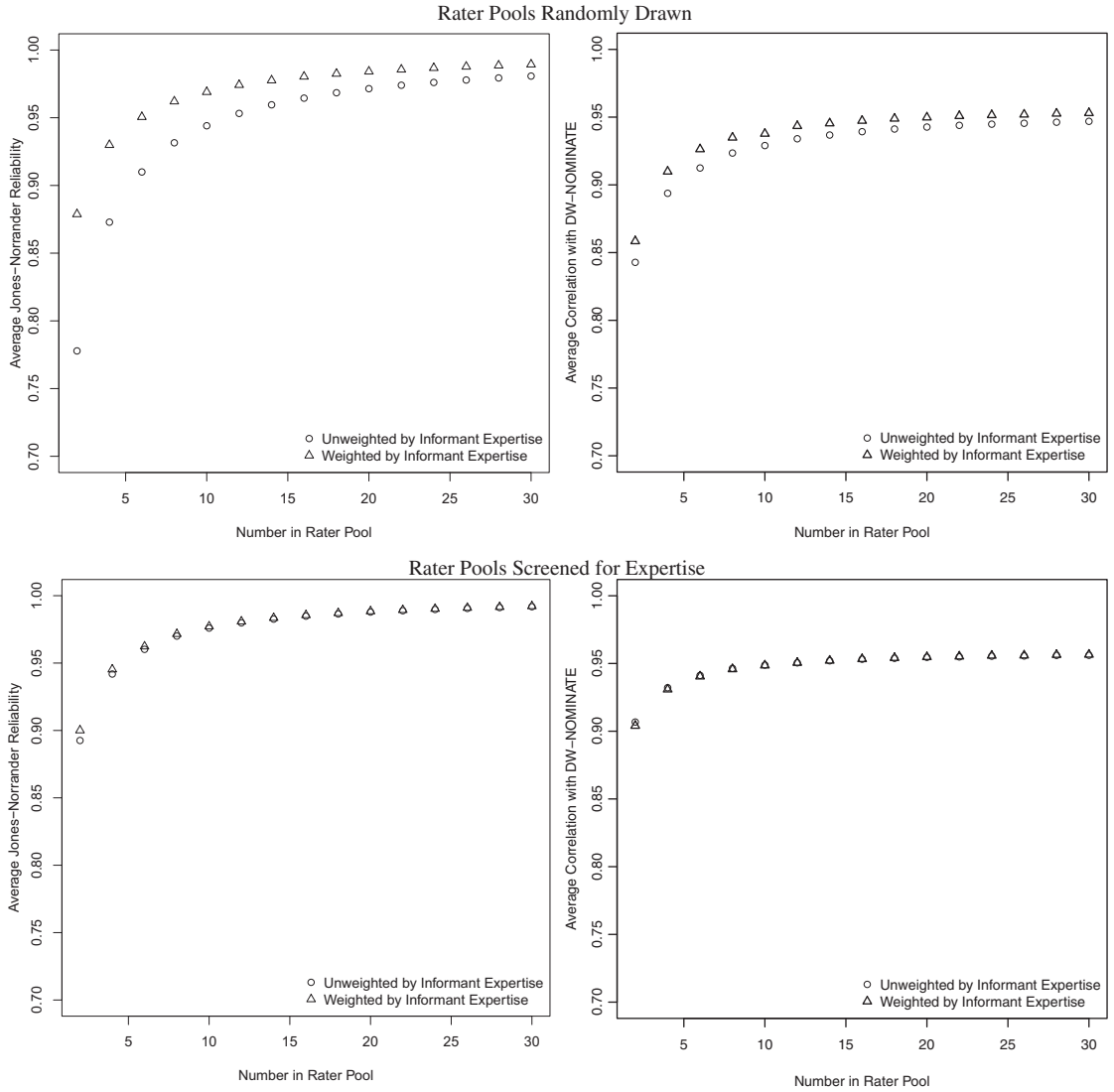

Fig. 5 Effects of expertise-based weighting and rater expertise on reliability and validity of aggregate incumbent position measures. Each point represents the mean reliability or correlation with DW-NOMINATE of aggregate measures based on one thousand samples, with replacement, from CCES respondents resident in a fixed set of districts (N of districts = 116).

therefore, lack clear criterion variables. In congressional elections, concepts related to candidate qualities, incumbent prospects, district context, and the like have proven difficult to measure and, as a result, are often excluded from models of candidate entry and election outcomes (but see Stone et al. 2010). Both the Candidate Emergence Study and the UCD-CES fill this gap through the use of informants to assess such qualities in congressional elections.[17] In this section, we focus on one variable of interest to congressional election studies—incumbents' prospects of winning re-election—and we use it along with the informant-based ideology (reconstructed as ideological extremity) to help explain incumbents' 2010 vote shares.

Scholars have appreciated for some time the importance of the strategic choices that incumbents and potential challengers make in anticipation of the election. These calculations are based, in part,

[17]See Stone et al. (2010) for analyses that incorporate informant-based measures of candidate prospects and valence, and Buttice and Stone (2012) for models with informant-based measures of candidate issue positions and valence.

on estimates of incumbent prospects, an estimate of how well the incumbent is likely to do in the upcoming election. Most studies use previous election outcomes, marginality, and the like as proxies for prospects but these fail to fully capture the concept. An important reason for the 2009 UCD-CES baseline informant survey prior to the 2010 general election study was to solicit informants' estimates of incumbent prospects well in advance of the electoral cycle, before most incumbents declare their intentions about whether to run for re-election and when potential challengers are estimating their chances of winning the seat.[18]

We compute four different versions of our informant-based measures of incumbent prospects and ideology for the Delegate and YouGov informant pools: a simple unweighted, unpurged mean; a mean of the informant judgments purged of individual-level partisan bias as described above; and two versions of the expertise-weighted means, also purged of partisan bias. This permits us to compare the performance of these measures built from a small, highly expert pool of informants (the Delegate pool) and moderate-sized, moderately expert pools of informants (the YouGov pool).

The calculation of the first two measures, the mean and the purged mean, is straightforward. Calculation of the two measures weighted for expertise requires more detailed description. The UCD-CES study built into the design a set of "expertise-checks" which we can use to construct expertise weights when aggregating the informant observations. The expertise checks consisted of knowledge questions about local political outcomes and actors, which can be compared to "true" outcomes in order to measure the expertise of informants: ability to place the two senators' ideology, predicted vote share of the 2010 winning candidate, and assessments of the 2008 presidential vote share in their district. For each item, we regressed the respective criterion variable on informant responses and calculate the squared residual for each informant. We standardize the squared residuals to have a mean of 0 and standard deviation of 1 and generated an expertise index by averaging across the items (alpha of 0.41 for the 2009 items and 0.47 for the 2010 items). We transformed the expertise variables to range from 1 to 100.

One shortcoming of studies advocating expertise weighting is the lack of guidance on appropriate values for α, the expertise weight parameter (Wagner, Rau, and Lindemann 2010). Because the effect of α is a function of both the choice of expertise scale and the distribution of respondents within each district on the scale, researchers are forced to experiment with potential values of α and scaling of expertise. We explored a number of different values for α ranging from 0 to 10 and found little variation in observed effects beyond an α value of 4 so we examine the case of weighting when α is 2 and when α is 4.

We ran OLS models of 2010 incumbent vote share in 127 House districts with complete data from two or more informant-respondents. The average number of informants per district used to create the incumbent prospects measures was 5.9 for the Delegate pool and 23.5 for the YouGov pool, and for ideological extremity, 4.3 and 23.5, respectively. Prior simulations in this article demonstrated greater reliability and validity of target-unit measures drawn from larger informant pools even when the larger informant pool is less expert. Thus, we expect to see stronger effects of our target-unit variables in the models that utilize the measures built from the YouGov informant pools. However, the substantive expectations for the informant measures in both models are that incumbents are penalized for ideological extremity, and that prospects should positively relate to vote share.

The statistical models also include standard predictors of incumbent vote share: vote share from the previous midterm election of the party that held the seat in the 2010 elections, an indicator of the partisan makeup of the district, the logged spending differential between the candidates, and

[18]The informant studies include a variety of questions related to anticipated electoral outcomes in the district, including the chances the incumbent will be challenged by strong opponents in the primary and general election stages, the chances the incumbent would run for re-election, and the chances the incumbent will win the primary and general election. The incumbent prospects variable is based on the incumbent general election chances item: "Please give your best estimate of how likely the incumbent in your district will win the general election if he/she wins the primary." Responses were on a seven-point scale ranging from "extremely unlikely" to "extremely likely."

Table 5 Informant-based measures and explaining incumbent vote share

	Aggregation strategy for informant data							
	Unweighted, unpurged mean		*Bias-purged mean*		*Weighted/purged mean (alpha = 2)*		*Weighted/purged mean (alpha = 4)*	
Delegate pool								
Incumbent ideological extremity	−0.66	(0.41)	−1.11**	(0.43)	−1.17**	(0.44)	−1.15**	(0.44)
Incumbent prospects	1.01**	(0.43)	1.10**	(0.42)	1.17**	(0.43)	1.18**	(0.43)
Incumbent vote share in 2006	0.11**	(0.03)	0.10**	(0.03)	0.10**	(0.03)	0.10**	(0.03)
Presidential vote share in 2008	0.68**	(0.05)	0.71**	(0.05)	0.71**	(0.05)	0.71**	(0.05)
Spending differential	1.74**	(0.32)	1.64**	(0.32)	1.58**	(0.32)	1.57**	(0.33)
Republican incumbent	11.29**	(1.02)	11.98**	(1.03)	12.16**	(1.07)	12.14**	(1.08)
Constant	6.05**	(2.70)	5.96**	(2.77)	6.09**	(2.79)	6.05**	(2.70)
R^2	0.92		0.92		0.92		0.92	
Observations	127		127		127		127	
YouGov pool								
Incumbent ideological extremity	−2.00**	(0.35)	−1.97**	(0.42)	−2.05**	(0.42)	−2.09**	(0.43)
Incumbent prospects	1.94**	(0.45)	1.77**	(0.45)	1.75**	(0.45)	1.74**	(0.45)
Incumbent vote share in 2006	0.06**	(0.03)	0.05*	(0.03)	0.05*	(0.03)	0.05*	(0.03)
Presidential vote share in 2008	0.72**	(0.04)	0.76**	(0.05)	0.77**	(0.05)	0.77**	(0.05)
Spending differential	1.45**	(0.29)	1.44**	(0.30)	1.48**	(0.30)	1.50**	(0.30)
Republican incumbent	11.86**	(0.93)	12.41**	(1.01)	12.44**	(1.00)	12.40**	(0.99)
Constant	8.74**	(2.60)	6.16**	(2.83)	5.83**	(2.84)	5.72**	(2.84)
R^2	0.94		0.94		0.94		0.94	
Observations	127		127		127		127	

*$p < 0.10$, **$p < 0.05$

Note. OLS regression coefficients and standard errors in parentheses. The dependent variable is the incumbent's two-party vote share. The different measures of incumbent ideological extremity and prospects have been rescaled to have a mean of 0 and standard deviation of 1 to facilitate comparisons across models.

whether the Republicans held the seat before the elections (to capture the national tide in favor of the GOP in 1910).[19]

Table 5 shows that these expectations are realized: extremist incumbents suffer a loss in vote share, ceteris paribus, and informants' judgments of incumbent prospects have independent effects on vote share. This is striking evidence that aggregated informant ratings can provide information not otherwise available using either approach. Lagged vote share, partisan makeup of the district, and party of the incumbent are standard indicators of incumbent prospects, yet the informant-based measure of prospects has a significant independent effect on vote share.

At the same time, we do see differences in the size and effects of variables built from the smaller Delegate pool compared to those built from the larger YouGov pool. Regardless of how the measure is calculated, in all cases, the coefficients for the YouGov-based measures are appreciably larger than the coefficients for the Delegate-based measures. The larger pool of informants in the YouGov survey provides more precise estimates of the incumbent quantities of interest, leading to more reliable and valid ordering of incumbents on the prospects scale compared to the smaller Delegate pools.

We see few substantive differences arise from the choice of post-survey data processing. The strength of the effects of ideological extremity in both models generally increases as we move from a simple mean to the more complex weighted and purged mean. Indeed, for the Delegate sample, the coefficient estimation on the unpurged mean is statistically insignificant. This is not entirely

[19]Previous literature explaining incumbent vote share offers much more sophisticated and nuanced models than we present here, particularly with regard to the role of incumbent prospects and district conditions. The model we present is admittedly simplistic, and therefore should only be interpreted as illustrating differences in performance of the various versions of the informant-based measures.

surprising given the small informant pools in some districts. Partisan biases of a single respondent can exert leverage over the target-unit-level measure in districts with two or three informants, and therefore, the partisan bias correction makes a substantial difference to the performance of the measure. This type of correction matters less for measures built from larger informant pools. In fact, we see larger proportional gains in the size of the effect of the Delegate-based measures, as our simulations predicted for ideology. We note a slight attenuation of the effect of incumbent prospects built from YouGov for the measures with post-survey bias correction and expertise weighting. It is possible that correcting the measure for bias leads to less variation in the prospects since most incumbents have very high prospects of winning. However, we do not see a similar attenuation in the model using the Delegate-based measures.

Although more complete models of incumbent vote share based on informant measures are possible, these results suggest the utility of these measures. The comparisons are also consistent with our conclusions in this article: larger YouGov informant pools generate measures that have consistently stronger effects than those built from the Delegate study. Purging for bias and weighting has relatively modest effects, especially in the YouGov studies, where the wisdom-of-crowds effect is most apparent.

5 Conclusions

Our results speak to questions about the design and manipulation of data for informant-based studies and highlight the value of using district observers to report attributes of US House candidates. Beyond ideology, candidate attributes such as their prospects of winning, their traits or personal qualities, and their skills as campaigners are important to understanding outcomes in congressional elections. The advantage of informant-based measures is that they open substantive questions in the study of voting and elections that heretofore have been difficult to broach because of a lack of systematic, comparable measures of challengers and incumbents. The method we apply to congressional elections should easily generalize to other types of measures and contexts. For example, expert surveys of party or elite positions could be augmented with observations from highly knowledgable individuals situated in the appropriate context. The penetration of online survey research firms into many industrialized countries makes the strategies used in this article increasingly viable and cheap, even outside the US context. The challenges and trade-offs in the design would likely be quite similar to those we have shown when estimating attributes of House incumbents because the fundamental rating task is similar.

We also suggest that political scientists could expand the use of informant-based studies to tap organizational data or local contextual data that are typically out of reach. In other fields, expert-based measures have been used to gauge risk and uncertainty related to civil infrastructure (Cooke and Goossens 2004), estimate species population (Martin et al. 2012), create indices of historical societal stressors (McCann 1998), and measure organizational and management characteristics of businesses (Phillips 1981; Boyer and Verma 2000; VanBruggen Lillien and Kacker 2002). It is easy to imagine a wide range of measures of interest to political science that draw from observations of organizational behavior or local political contexts.

We have seen that moving from relatively small pools of informant respondents to modestly larger pools per unit can substantially improve the reliability and validity of measures. For example, informant pools of approximately 25 appear to be large enough to ameliorate the effects of relying on informants of lower average expertise, compared with small pools of high-expertise informants. In the case of opt-in panel studies, researchers should consider prescreening respondents for expertise. We recommend this as a further hedge, but we saw clear evidence that lower-expertise individual ratings can be superior to high-expertise ratings with a modest increase in N. When employing mass-respondent informants who have not been prescreened for expertise, post-selection weighting can have a modest beneficial effect, as can correcting for individual partisan bias.

Survey methods are evolving as researchers strive to keep costs down, manage high nonresponse levels, and incorporate experimental and other types of variability into their designs. Opt-in panels can be a useful part of the survey researcher's arsenal in confronting these sorts of issues. A side

benefit of the growth in opt-in panels is that they provide a huge reservoir of potential informants on a wide range of topics of interest to social scientists. Traditionally, researchers have thought of surveys as primarily devices for soliciting "expert responses" about the self. There is every reason to expand our reach to solicit responses about targets external to the self, especially when the benefits of aggregation can be harnessed to reduce the effects of individual errors. We suspect our experience in tapping unmeasured dimensions in the study of congressional elections could be replicated by scholars interested in these phenomena and more. If so, the accumulation of scientific knowledge about politics will advance to a broader set of frontiers as scholars recognize the benefits of this approach.

Funding

Data for portions of this article were funded by the National Science Foundation Grant [SES-0852387, W.J.S., P.I.] "Political Context and Citizen Response in the 2010 Elections."

References

Albright, Jeremy J., and Peter Mair. 2011. Does the number of parties to place affect the placement of parties? Results from an expert survey experiment. *Electoral Studies* 30(4):858–64.

Andersson, Patric, Jan Edman, and Mattais Ekman. 2005. Predicting the World Cup 2002: Performance and confidence of experts and non-experts. *International Journal of Forecasting* 21(3):565–76.

Ansolabehere, Stephen. 2010. CCES Common Content, 2010. http://hdl.handle.net/1902.1/17705 V3 [Version] (accessed January 18, 2014).

Benoit, Kenneth, and Michael Laver. 2006. *Party policy in modern democracies*. London: Routledge.

Boyer, Kenneth K., and Rohit Verma. 2000. Multiple raters in survey-based operations management research: A review and tutorial. *Production and Operations Management* 9(2):128–40.

Braouezec, Yann. 2010. Committee, expert advice, and the weighted majority algorithm: An application to the pricing decision of a monopolist. *Computational Economics* 35:245–67.

Budge, Ian. 2000. Expert judgments of party policy positions: Uses and limitations in political research. *European Journal of Political Research* 37:103–13.

Buttice, Matthew K., and Walter J. Stone. 2012. Candidates matter: Policy and quality differences in congressional elections. *Journal of Politics* 74(3):870–87.

Carroll, Royce, Jeffrey B. Lewis, James Lo, Keith T. Poole, and Howard Rosenthal. 2009. Measuring bias and uncertainty in DW-NOMINATE ideal point estimates via the parametric bootstrap. *Political Analysis* 17(3):261–75.

Castles, Francis G., and Peter Mair. 1984. Left-right political scales, some "experts'" judgments. *European Journal of Political Research* 12(1):73–88.

Clinton, Joshua D., and David E. Lewis. 2008. Expert opinion, agency characteristics, and agency preferences. *Political Analysis* 16(1):3–20.

Cooke, Roger M., and Louis H. J. Goossens. 2004. Expert judgement elicitation for risk assessments of critical infrastructures. *Journal of Risk Research* 7(6):643–56.

Dancy, Logan, and Geoffrey Sheagley. 2013. Heuristics behaving badly: Party cues and voter knowledge. *American Journal of Political Science* 57(2):312–25.

Feldman, Stanley, and Pamela Conover. 1983. Candidates, issues and voters: The role of inference in political perception. *Journal of Politics* 45(4):810–39.

Gaissmaier, Wolfgang, and Julian N. Marewski. 2011. Forecasting elections with mere recognition from small lousy samples: A comparison of collective recognition, wisdom of crowds, and representative polls. *Judgment and Decision Making* 6(1):73–88.

Hooghe, Liesbet, Ryan Bakker, Anna Brigevich, Catherine De Vries, Erica Edwards, Gary Marks, Jan Rovny, Marco Steenbergen, and Milada Vachudova. 2010. Reliability and validity of the 2002 and 2006 Chapel Hill Expert Surveys on party positioning. *European Journal of Political Research* 49:687–703.

Huber, John, and Richard Inglehart. 1995. Expert interpretations of party space and party locations in 42 societies. *Party Politics* 1(1):73–111.

Huckfeldt, Robert, Edward G. Carmines, Jeffery J. Mondak, and Eric Zeemering. 2007. Information, activation, and electoral competition in the 2002 congressional elections. *Journal of Politics* 69(3):798–812.

Javaras, Kristin N., H. Hill Godlsmith, and Nan M. Laird. 2011. Estimating the effect of a predictor measured by two informants on a continuous outcome: A comparison of methods. *Epidemiology* 22(3):390–99.

Jones, Bradford S., and Barbara Norrander. 1996. The reliability of aggregated public opinion measures. *American Journal of Political Science* 40(2):295–309.

Kitschelt, Herbert, and Daniel M. Kselman. 2012. Economic development, democratic experience, and political parties linkage strategies. *Comparative Political Studies* September:1–32.

Lorenz, Jan, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing. 2011. How social influence can undermine the wisdom of crowds effect. *Proceedings of the National Acadamy of Science* 108(22):9020–25.

Maestas, Cherie D., Matthew K. Buttice, and Walter J. Stone. 2013. Replication data for: Extracting wisdom from experts and small crowds: Strategies for improving informant-based measures of political concepts. http://dx.doi.org/10.7910/DVN/23170 IQSS Dataverse Network [Distributor] V1 [Version] (accessed January 18, 2014).

Martin, Tara G., Mark A. Bergman, Fiona Fidler, Petra M. Kuhnert, Samantha Low-Choy, Marissa McBride, and Kerrie Mengersen. 2012. Eliciting expert knowledge in conservation science. *Conservation Biology* 26(1):29–38.

McCann, Stewart J. H. 1998. The extended American social, economic, and political threat index (1788–1992). *Journal of Psychology* 132(4):435–49.

Page Scott, E. 2007. *The difference: How the power of diversity creates better groups, firms, schools, and societies*. Princeton, NJ: Princeton University Press.

Phillips, Lynn W. 1981. Assessing measurement error in key informant reports: A methodological note on organizational analysis in marketing. *Journal of Marketing Research* 18(4):395–415.

Poole, Keith T., and Howard Rosenthal. 1997. *Congress: A political-economic history of roll call voting*. New York: Oxford University Press.

Powell, Lynda. 1989. Analyzing misinformation: Perceptions of congressional candidates' ideologies. *American Journal of Political Science* 33:272–93.

Saiegh, Sebastian M. 2009. Recovering a basic space from elite surveys: Evidence from Latin America. *Legislative Studies Quarterly* 34(1):117–45.

Sjoberg, Lennart. 2009. Are all crowds equally wise? A comparison of political election forecasts by experts and the public. *Journal of Forecasting* 28(1):1–18.

Steenbergen, Marco, and Gary Marks. 2007. Evaluating expert judgments. *European Journal of Political Research* 46:347–66.

Stone, Walter J., and Elizabeth N. Simas. 2010. Candidate valence and ideological positions in U.S. House elections. *American Journal of Political Science* 54(2):371–88.

Stone, Walter J., Sarah A. Fulton, Cherie D. Maestas, and L. Sandy Maisel. 2010. Incumbency reconsidered: Prospects, strategic retirement, and incumbent quality in U.S. House elections. *Journal of Politics* 72(1):178–90.

Stone, Walter J., L. Sandy Maisel, and Cherie D. Maestas. 2004. Quality counts: Extending the strategic politician model of incumbent deterrence. *American Journal of Political Science* 48(3):479–95.

Surowiecki, Jame. 2004. *The wisdom of crowds*. New York: Random House.

Van Bruggen, Gerrit H., Gary L. Lilien, and Manish Kacker. 2002. Informants in organizational marketing research: Why use multiple informants and how to aggregate responses. *Journal of Marketing Research* 39(4):469–78.

Wagner, Stephan M., Christian Rau, and Eckhard Lindemann. 2010. Multiple informant methodology: A critical review and recommendations. *Sociological Methods and Research* 38(4):582–618.

Whitefield, Stephen, Matilda Anna Vachudova, Marco R. Steenbergen, Robert Rohrschneider, Gary Marks, Matthew P. Loveless, and Liesbet Hooghe. 2007. Do expert surveys produce consistent estimates of party stances on European integration? Comparing expert surveys in the difficult case of Central and Eastern Europe. *Electoral Studies* 26(1):50–61.