

Assessing Political Knowledge: Problems and Solutions in Online Surveys

Douglas J. Ahler
douglas.j.ahler@vanderbilt.edu
Vanderbilt University

Stephen N. Goggin
stephen.goggin@sdsu.edu
San Diego State University

May 2, 2017

Abstract

Despite their ubiquity in online survey research, political knowledge (PK) measures are often constructed haphazardly, in many cases using items and batteries designed decades ago. We highlight four concerns: 1) threats to validity from limited temporal and topic variation, 2) false negatives stemming from online respondents' higher-than-average knowledge, 3) item formatting choices that limit comparability and offer "researcher degrees of freedom," and 4) reliability and cost tradeoffs. Researchers can address these concerns by building novel PK batteries with specific item and distributional properties in mind, attending to item format, and tailoring item content to the study at hand. Using evidence from three separate studies with a wide variety of political knowledge items and formats, as well as item response theory (IRT) models and Monte Carlo simulations, we demonstrate the advantages of our approach. These easily-implemented strategies make political knowledge measures more valid, cost-efficient, and useful as moderators in experiments.

Assessing citizens' knowledge of public affairs is among the longest-standing research agendas in the study of political behavior (e.g., Barabas et al. 2014; Lippman 1922; Delli Carpini and Keeter 1996; Lupia 2016). Scholars revisit this topic primarily because of its normative implications. Citizens' political knowledge—or lack thereof—spurs important questions about civic competence, the quality of the political information environment, and, ultimately, democratic accountability. For example, citizens with high political knowledge are more likely to be tolerant of opposing viewpoints (Delli Carpini and Keeter 1996; Galston 2001), consistent in their policy attitudes (Converse 1964), and to make voting decisions that comport with their attitudes (Althaus 1998; Bartels 1996; Delli Carpini and Keeter 1996), all of which are generally seen as democratic “goods.” Conversely, citizens with low political knowledge are more likely to choose candidates based on their appearance (Ahler et al. 2017; Lenz and Lawson 2011) and to believe falsities in the information environment (Nyhan and Reifler 2010, c.f., Roush 2016). Because political knowledge is so fundamental, political scientists employ it in diverse applications: in a descriptive, univariate manner (e.g., “How much does the ordinary American know, and how has that changed over time?”), as an independent variable (e.g., “How does political knowledge affect attitudinal constraint?”), as a dependent variable (e.g., “Does civics education improve political knowledge?”), and, perhaps most often, as a moderating variable (e.g., “How much more does a candidate’s appearance influence less-informed citizens than the most informed?”). “Political knowledge” thus returns nearly 16,000 Google Scholar hits for articles published in the past decade (2008-2017)—including 66 hits in the *American Political Science Review*, 92 in the *American Journal of Political Science*, and 331 in the *Journal of Politics*.

However central its place, political knowledge is often poorly measured. Researchers often construct political knowledge batteries in an impromptu and slapdash manner, usually failing to justify design choices and ignoring basic lessons from the educational testing literature (Delli Carpini and Keeter 1996; Lupia 2016). Such “researcher degrees of freedom” may yield false positives between political knowledge and other variables of interest (Lupia 2016).

On the other hand, as we show, ill-conceived batteries can also obscure real, meaningful relationships. In particular, batteries that are too easy or difficult yield skewed distributions and reduced variation in scores, which in turn attenuate observed relationships between political knowledge and other variables.

This problem is exacerbated by several contemporary challenges. The first is theoretical. Recent work by Barabas et al. (2014) demonstrates that political knowledge is not one monolithic concept. Despite this advance, it is still almost always conceptualized as such at the research design stage. This may lead to slippage between concepts and operationalized measures, and thus attenuation bias in observed relationships (Adcock and Collier 2001).

A second problem is practical. Researchers still routinely measure political knowledge with batteries developed decades ago (most notably, Delli Carpini and Keeter 1996) with ad hoc modifications to fit new survey modes. However, survey sampling, administration, and educational testing have changed dramatically since their development. Most notably, batteries designed pre-2000 tend to be optimized for phone and in-person surveys, and for reasons we elaborate below, are liable to be excessively easy. They can thus yield attenuation bias in observed relationships when deployed in web surveys.

In this paper we propose a two-pronged solution to address these theoretical and pragmatic concerns. First, we argue that researchers should update the items in political knowledge scales regularly and tailor them to the topical focus of the study at hand. By doing so, researchers mitigate problems of panel conditioning and are more likely to operationalize political knowledge appropriately for particular applications. Second, we suggest the use of the true/false item format to minimize researcher degrees of freedom, reduce survey time and cost, and increase inter-item comparability across topics. We demonstrate the advantages of these approaches with three studies on three separate online survey samples, highlighting how they minimize the concerns we have detailed.

Desirable Properties of Political Knowledge Batteries

As with all measures, we would like our instruments for assessing political knowledge to be reliable, precise, and valid. Because context, and not just random error, may render individual items more or less difficult, reliability is usually assessed *inter-item*.¹ That is, if we believe that the items in a battery all measure the same underlying construct—in this case, political knowledge—then our battery is more reliable when individual items are measured with minimal random error and, thus, correlate more highly (Hoyle, Harris and Judd 2002). However, we also note that political knowledge is a multidimensional concept, meaning that short scales may suffer from lower inter-item reliability if they try to cover multiple dimensions (Barabas et al. 2014). Thus, a short scale’s reliability may decrease simply due to broader topical coverage.

Educational testing theory conceptualizes this potential tradeoff between reliability and *precision*, the relative fineness or coarseness of one’s measure. Generally, a measure becomes more reliable as the number of scale points increases—until measures become more precise than the concept is in respondents’ minds, at which point reliability begins to decline (Nunnally and Bernstein 1994). However, political knowledge is an interesting case. Since it is usually measured indirectly by administering factual questions about politics, precision-in-measures should ideally match the conceptual precision in the *researcher’s* mind. Since we tend to conceive of political knowledge as a continuous concept, reliability is unlikely to decline with additional items.²

On the other hand, there are diminishing marginal returns to reliability for each additional item added to the battery (Nunnally and Bernstein 1994)—and opportunity costs to long

¹For an example of context effects, a newsworthy Supreme Court case may temporarily increase survey respondents’ ability to identify the Chief Justice.

²Notably, however, depending on the dimensionality of political knowledge, a change in the composition of items might yield less-reliable sub-scales.

political knowledge batteries. Survey research is often priced by time or number of items. Thus, the ideal political knowledge battery should maximize reliability and precision within the project’s constraints. Ergo, if the researcher is paying by the minute, she should maximize the number of PK items administered in a set amount of time; this may have implications for the format one uses for PK items. If the researcher is paying by the item, our suggestions below about building a battery with ideal distributional properties will be most helpful. In sum, we want our measures not only to be reliable and precise, but also efficient.

Finally, like all measures, it is of paramount importance that political knowledge batteries are *valid*—that they reflect the underlying concept of interest. Indeed, we tend to measure political knowledge with factual tests because, “more directly than any alternative measures, (these batteries) capture what has actually gotten into people’s minds, which, in turn is critical for intellectual engagement with politics” (Zaller 1992, p. 21). However, as we explain further in the next section, recent research challenges the notion that political knowledge is one monolithic concept (Barabas et al. 2014), raising new questions about valid measurement.

While battery scores are often seen as mere proxies for actual political knowledge (e.g., Lupia 2006; Zaller 1992)—making conventional validation exercises difficult—we can assess various batteries’ appropriateness through construct validation. Political science has documented and replicated several findings related to political knowledge—e.g., its associations with participation (Galston 2001) and attitudinal constraint (Converse 1964). By assessing how clearly these relationships appear with new measures of political knowledge, we can compare new measures’ performance to established baselines. (Cronbach and Meehl 1955).

In addition to these basic measurement properties, there are particular distributional properties researchers ought to strive for when administering political knowledge batteries—especially because political knowledge is so frequently hypothesized to moderate relationships of interest. When investigating these moderating effects, researchers ask, “How, if at all, does the relationship between X and Y differ at one level of political knowledge compared to another?” In doing so, researchers are juxtaposing one subset of respondents, defined by

their level of political knowledge, against another. Comparisons like these require statistical power to reject null hypotheses, and statistical power is maximized when observations are distributed evenly into strata (Thompson 2012). The two-strata case is clearest: when comparing a high-knowledge group to a low-knowledge group, the “effective n ” is the size of the smaller group, so power is maximized when respondents divide evenly into groups. However, this logic generalizes to more finely-defined groups. As such, if we conceptualize political knowledge as continuous, then the ideal distribution of scores on a PK battery would approximate the discrete uniform distribution, which features the desirable property: $n_{[b,c]} = n_{[b+i,c+i]} \forall b, c, i$ s.t. $b, c, (b+i), \& (c+i) \in [a, d]$.

According to similar logic, a distribution that covers the full range of possible scores is preferable to a distribution that covers only a partial range. A second desirable property of a discrete uniform distribution is its high variance of observed scores, implying sufficient data points to compare across the range of observations.³ All else equal with score distributions (and latent knowledge distributions), a truncated range of observed scores implies lower variance, which in turn implies reduced power for detecting meaningful associations. We can also think about this problem in terms of precision: if we administer a 10-item battery and the lowest score is 3, then we have effectively administered a scale with eight points instead of eleven and failed to maximize sensitivity within the constraints of the research design.

In sum, researchers should try to achieve distributions that are relatively symmetrical, with high variance and with relatively consistent density across the range of scores. We note these desirable properties for two reasons. First, political knowledge batteries are all-too-

³Indeed, the discrete uniform distribution maximizes the possible variance of a discrete random variable X with a bounded range. The variance of a discrete uniform distribution is given by $\frac{1}{12}((b-a+1)^2 - 1)$. The maximum variance of any random variable X is given by $E(X) \cdot E(1-X)$. The latter is maximized when $E(X) = E(1-X)$, a property the uniform distribution satisfies by virtue of being the maximum entropy probability distribution for discrete distributions bounded by the interval $[a, b]$.

often constructed on an impromptu basis to assess questions of moderation and mediation. These types of complex hypotheses often face significant power challenges to begin with. Failure to consider the distribution of the moderating/mediating variable only adds to the likelihood of Type II errors. Second, however, we raise these points to demonstrate that researchers have some control over the distribution of these variables. Political knowledge scores reflect respondents' answers to factual questions, and researchers can thus build batteries to achieve not only validity, reliability, and precision in measures, but also to achieve distributions of scores that maximize statistical power.

In recent work, many researchers have sought to address concerns regarding the distributional properties of political knowledge (and other scaled concepts) by using Item Response Theory (IRT) models to estimate the underlying hypothesized latent distribution. With an item response function, researchers may estimate each item's difficulty and discrimination (the degree to which the individual item predicts respondent knowledge), and in three-parameter models, a pseudo-guessing parameter (described in greater detail in Footnote 13). By scaling respondents and items accordingly, IRT models thus may help "correct" a response distribution, effectively re-weighting items into a more normal distribution. However, we eschew the need for corrective IRT procedures. While they can help fix "broken" scales, we argue the goal should be to design a scale with desirable properties up front such that correction is not necessary. If a scale is well-designed, we should expect IRT-scaled and additively-scaled measures to be distributed quite similarly.

Contemporary Challenges

Theoretical and Conceptual Issues

Valid measurement is paramount not just for political knowledge but for all concepts. Traditionally, however, because political knowledge scales have been seen as "proxy measures" for media attention, political interest, political sophistication, etc. (e.g., Brewer 2003;

Zaller 1992), they have not faced the tough questions about validity that other common scales have. We believe that this *carte blanche* assumption is unwarranted, especially in light of recent research suggesting that political knowledge is more conceptually complex than a monolithic store of facts related to public affairs.

In particular, Barabas et al. (2014) identify and demonstrate the theoretical relevance of two distinct dimensions of political knowledge, implying that there exist at least four distinct types of political knowledge questions. The research suggests that questions vary on a temporal dimension; that is, political facts may be long-standing or instead hinge on recent events. Importantly, education appears to affect knowledge of time-invariant facts, while attention to mass media correlates more strongly with knowledge of temporally-specific facts. This suggests that researchers ought to think more critically about what they are using political knowledge as a proxy for. For example, when attempting to construct a proxy measure of media attentiveness, instead of simply pulling a battery off the proverbial shelf, researchers may find greater validity in building batteries from items drawn from recent news stories.⁴ Conversely, a battery composed entirely of current events questions lacks face validity as a proxy for static civics knowledge.

In addition to this temporal dimension, political knowledge appears to vary on a topical dimension. Battery items can engage either general political awareness or policy-specific knowledge (Barabas et al. 2014). Again, researchers can improve their PK measures' face validity by selecting items that properly tap the concept they want to measure. This suggestion is not wholly novel; it has merely gone unheeded. As Converse (1975, p. 79) describes,

“Many studies assess information levels, leaning inevitably toward rather stray facts as the basis of the “test,” and then quite independently attempt to measure pure opinions on meatier dimensions of public debate. What is rarely done is to

⁴The classic five-item (Delli Carpini and Keeter 1996) battery features two static civics items, two time-variant recall items, and one item that is more difficult to classify on the Barabas et al. (2014) static-surveillance dimension (party placement).

explore the information base underlying the opinions themselves. But when this is done the results are frequently both amazing and instructive.”

Relating knowledge measures to the study at hand is especially important for properly identifying causal effects. For example, Lenz, Turney and Freeder (2016) provide evidence that specific knowledge of the parties’ positions on issues—rather than general political knowledge—explains Conversian attitude crystallization. In turn, this provides a more pessimistic explanation for correspondence between issue positions and candidate choice among the “knowledgeable” (e.g., Ansolabehere, Rodden and Snyder 2008): party following, not issue voting. Attention to the topical dimension also seems particularly important when researchers are interested in citizens’ knowledge of a particular policy domain. While a civics test easily fails face validation in this case, a general policy knowledge quiz does as well. The obvious solution in cases like these is to tailor batteries to the domain of interest, as it then more accurately measures the moderating role of knowledge in that domain. For example, if interested in knowledge of foreign affairs (e.g., Aldrich, Sullivan and Borgida 1989; Baum 2003; Miller and Stokes 1963), battery items should engage that topic.

Practical Issues Related to Contemporary Surveying

The gold-standard political knowledge battery (Delli Carpini and Keeter 1996) selected five items that performed optimally on the 1990-91 ANES. Since that time, survey research has changed dramatically. Most notably, surveys today are generally conducted over the web instead by phone (or in person, like the ANES). Political knowledge questions are usually administered as *recall* items over the phone or in person. By contrast, they are usually administered as *recognition* items on the web, as this allows for automated scoring.⁵ By virtue of priming and guessing, recognition items garner higher accuracy. Thus, Delli Carpini

⁵This strategy also precludes validation challenges with recall items (e.g., Mondak 2001). For example, how should we score an MTurker’s response that John Roberts is the “Chief of Justice”? While technically incorrect, such a response demonstrates more knowledge than a

and Keeter’s (1996) battery is almost certainly easier in most contemporary contexts than it is in the contexts for which it was designed.⁶

This problem is likely compounded by additional developments over the past two decades. Across all surveying modes, response rates have plummeted (Curtin, Presser and Singer 2005). Furthermore, with a few exceptions, even the highest-quality online samples tend to be purposive rather than probability-based. These issues raise questions about selection bias: contemporary respondents may be unusually interested in politics, and thus more knowledgeable (Krupnikov and Levine 2014). Respondents in online panels may also participate in scores, if not hundreds, of studies (Hillygus, Jackson and Young 2014), and may learn facts about politics they otherwise would not.⁷ In sum, political knowledge batteries that were optimally difficult in the past are likely to be too easy today. They are thus liable to yield score distributions which fail to separate middling-knowledge and high-knowledge respondents and which lack sufficient observations at the low end of the distribution.

blank or wholly inaccurate one. Similarly, recognition items allow researchers to dodge the question of expressive responses that seem to demonstrate some knowledge, such as another MTurk respondent who proclaimed John Roberts “Chief of (epithet omitted).”

⁶A seeming advantage of existing batteries is that the items themselves have remained relatively constant over time, allowing for longitudinal analyses. However, the variation in response format over time makes bridging across these items over time incredibly hard and requires a bevy of statistical assumptions. As such, there is little reason to prefer existing batteries for means of longitudinal comparisons, as these comparisons are already fraught.

⁷A related concern is that respondents in online panels may be highly motivated to respond correctly to items, thus cheating (Clifford and Jerit 2016).

Implementing Design-Based Solutions

We now describe our solution for addressing these concerns. While it does not completely render these concerns null, it does minimize them, and as we will demonstrate, produces a number of other advantages. Together, our measurement strategy addresses four key concerns with existing strategies:

1. Restricted domains of both temporal & topic variation
2. Higher-than-average knowledge & panel conditioning
3. Response options & item format limit comparability with guessing corrections
4. Reliability & time/cost tradeoffs

Our solution, while unconventional for most papers focused on measurement, is simple: researchers should design their own political knowledge scales that best suit the study at hand. In building these scales, we offer guidance on picking item topics, picking item content, constructing items, as well as scaling these items together optimally. By designing one's own battery, researchers inherently address concerns #1 and #2 above, and by standardizing question content and format as we describe, also address concerns #3 and #4. In the following sections, we describe how and why our strategy addresses they concerns, marshaling evidence from three separate experimental studies on three unique online survey platforms.

Empirical Studies

Study 1 was fielded in February 2016 on Amazon's Mechanical Turk ($N = 422$). In addition to questions about standard demographics and political attitudes, respondents completed a 42-item knowledge battery, spanning across six political topics (civics, people, mass politics, domestic politics, international politics, party placement on issues), and including six items on celebrity/entertainment knowledge for comparison purposes. Respondents were

randomly assigned to answer these questions in one of two formats: a four-response multiple choice format, or a true/false format.⁸ From this study, we can assess not only the impact of varying question format, but the relative difficulty of particular topics and questions. We use Item Response Theory (IRT) models to estimate item parameters for the full 36-item political knowledge scale and shorter subsets, as well as as a series of Monte Carlo simulations to assess the sensitivity of these scales to the inclusion/exclusion of particular items with varying properties.

Study 2 was fielded in July 2016 on a survey of California respondents ($N = 3252$, fielded via Survey Sampling International). In this study, we randomized respondents to either receive the canonical five-item Delli-Carpini & Keeter political knowledge scale, or a five-item scale built from optimal questions from our February 2016 study. Each item was formatted as a four-response multiple choice item, and additionally, respondents were randomized to either an “easy” or “hard” set of distractor items. This study allows us to examine the relative performance of a tailored scale versus the canonical measure, and also to demonstrate the sensitivity of PK scales to variation in distractor options when using a four-choice multiple choice format.

Finally, Study 3 was fielded in October 2016 on a survey of California respondents ($N = 1800$, fielded via YouGov). In this study, all respondents saw a six-item political knowledge scale. The difficulty of phrasing and response options were randomly assigned within each item. This study allows us to further assess not just how the difficulty of distractor options affects overall scale performance, but how phrasing choices unrelated to the item’s topical content can affect the item’s difficulty and discriminatory power.

⁸All exact question wordings for all studies can be seen in the Online Appendix (OA), sections OA-2-OA-4. For all items, response options were randomly rotated in order, except for cases in which a natural numeric ordering to response options exists.

Designing a Political Knowledge Battery

In this section, we describe how our approach addresses the preceding concerns in three steps. First, we highlight our recommended procedure for choosing item topics and how to choose content for specific items within these topics. Next, we demonstrate item difficulty’s sensitivity to the multiple-choice format, noting the relative advantage of true/false items. Finally, we discuss methods for scaling items together. Empirical evidence from all three studies is grouped together by topic, with each table/figure noting its source study.

Choosing Item Topics

Political knowledge is not monolithic (e.g., Barabas et al. 2014; Delli Carpini and Keeter 1996; Johnson 2009; Schudson 1998). People are not only heterogeneous in how much they know, but also what they know about. What people know and don’t know often map onto subcategories of more general “political knowledge.” As discussed above, individual PK items reflect both a topical dimension (general versus policy-specific information) and a temporal dimension (how recently this information came into being, and how long it is likely to remain a relevant fact, Barabas et al. 2014). So, for example, some citizens may be more likely to recall static facts about civics, while others may be more likely to recall current events related to social movements. In the aggregate, some subcategories of knowledge items may thus have greater difficulty than others.

If one’s goal is to maximize spread on a political knowledge battery across a variety of topics, a researcher requires information about the likely difficulty of potential items *for the sample to which the battery will be administered*. A vast array of previous research on online survey samples, particularly Amazon’s Mechanical Turk, has shown these respondents to be more knowledgeable and educated than representative samples of Americans (Berinsky, Huber and Lenz 2012). Thus, items designed for representative samples may in fact be too easy for online samples. As shown in Figure 1, the oft-used Delli Carpini & Keeter scale

results in an average score of 0.72 ($Var = 0.054$) on a 0-1 scale, while an updated battery (discussed in greater detail below) yields a lower average score of 0.58 ($Var = 0.066$) and a less skewed distribution across the range of scores.⁹

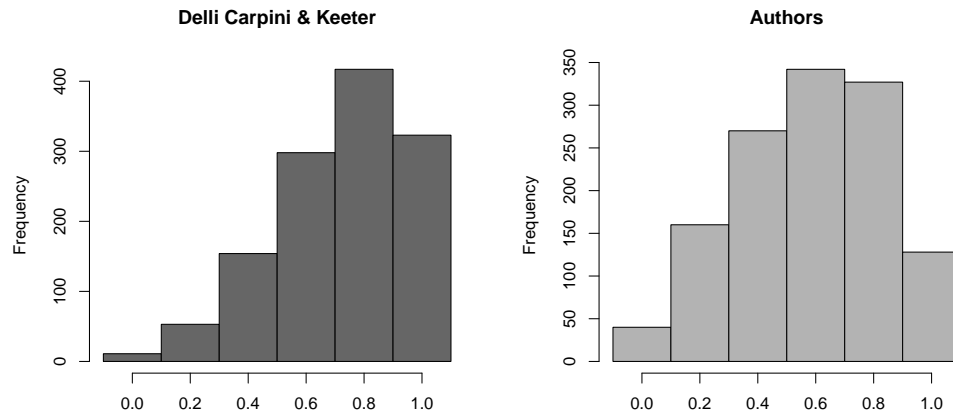


Figure 1: Scale difficulty in online sample [Study 2]
 NOTE: Mean for Delli Carpini & Keeter scale is 0.72, (Authors) is 0.58

What happens if we choose items that are too hard or too easy? Besides the obvious mean difference in score created by this problem, it attenuates any relationship with other variables of interest. Figure 2 shows this problem by displaying the results of a bivariate, locally-weighted regression of constraint, measured as how consistently one takes liberal positions on issues (scaled by IRT and then folded), on political knowledge—a well-documented association in public opinion (e.g., Bartle 2000; Delli Carpini and Keeter 1996; Luskin 1987). Each of the three lines represents a different regression with a five-item political knowledge scale constructed from our data. The orange line represents an optimally-constructed battery: it maximizes the within-battery standard deviation of scores among batteries that fall in the 45th through 55th percentile on average IRT difficulty parameter. (That is, it maximizes spread among batteries that are neither “too hard” nor “too easy” on average.) By contrast, the blue and green lines represent the hardest five-item battery and the easiest five-item battery, respectively. (See OA-5 for the items that went into each of these three

⁹Items are scaled together in a simple additive score, rescaled between 0-1.

batteries.) Notably, while we obtain the expected strong positive relationship for the average battery, we see a much weaker relationship for both the easy and hard batteries.¹⁰

The figure’s use of LOESS curves (rather than simple OLS slopes) helps to illuminate why this is the case. While the relationship between constraint and scores on the optimal battery is strong and closely approximates a linear relationship, we do not observe these properties for the suboptimal batteries. In particular, the relationship between scores on the too-hard battery and constraint flattens as scores near 0.5 (on a 0-1 scale). This is because relatively few respondents score highly on the too-hard battery; guessing likely accounts for differences between those scoring relatively highly, meaning that meaningful variation only exists on the lower half of the scale. Similarly, the relationship between scores on the too-easy battery and constraint is quite flat over the lower half of the knowledge scale.

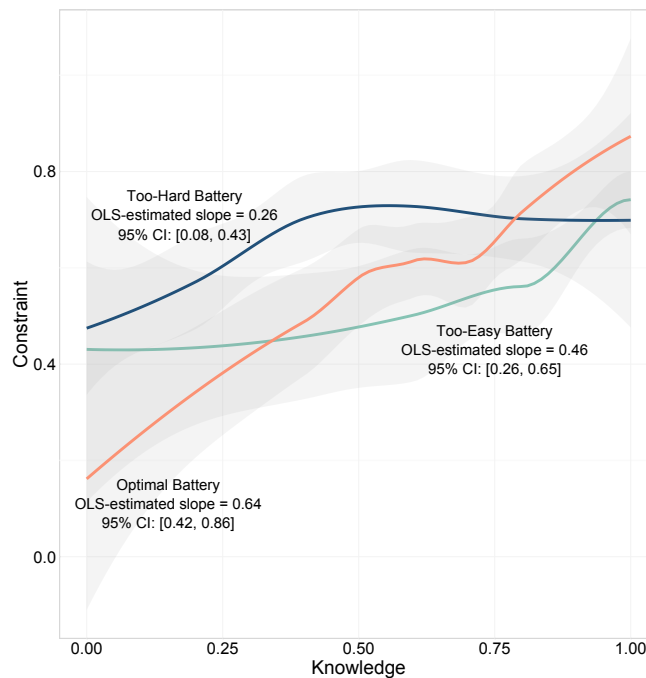


Figure 2: Poorly-Constructed Knowledge Batteries Underestimate the Relationship Between Knowledge and Constraint [Study 1]

NOTE: LOESS smoother with 95% confidence intervals depicted in gray.

¹⁰Scaled from 0-1, the optimal battery has mean = 0.68, $Var = 0.054$. The too-hard battery has mean = 0.44, $Var = 0.070$. The too-easy battery has mean = 0.85, $Var = 0.052$.

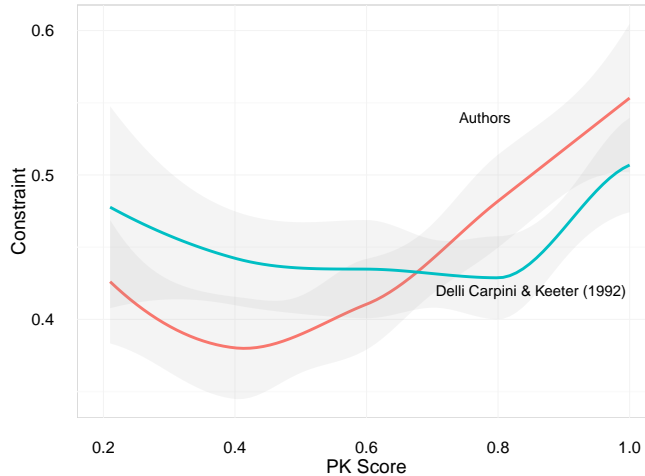


Figure 3: Skewed Knowledge Batteries Underestimate the Relationship Between Knowledge and Constraint [Study 2]

NOTE: LOESS smoother with 95% confidence intervals depicted in gray.

We can see this same pattern in Figure 3 when we compare the performance of the canonical scale with our newly adapted one. Over much of the distribution, we observe a relatively flat relationship between ideological constraint and political knowledge, measured with the Delli Carpini & Keeter battery. In particular, the canonical battery yields this flat relationship over the range of scores less than 0.8 on the 0-1 scale, implying that it fails to separate medium-information respondents from the truly tuned out—at least when administered on this particular online platform with recognition items. By contrast, our battery produces the expected positive relationship for much of the scale.¹¹ This may be simply because our items yield a more difficult scale, but panel conditioning may also account for the canonical battery’s failure in this context. Because many online survey respondents take political surveys on a regular basis, it is likely they have previously encountered “off-the-

¹¹That the relationship in Figure 3 is nonmonotonic is not inconsistent with classic studies of constraint. Citizens at the low end of the knowledge distribution may be constrained for non-policy reasons (Converse 1964). For example, low-information partisans may report constrained opinions via party-following (Lenz 2012), while independents who vote purely based on the “nature of the times” or group-interest voters may know more but demonstrate less opinion constraint.

shelf” political knowledge items. In addition to producing a more difficult scale, our solution also helps to preclude artificially high scores from respondent familiarity with particular items.

The Solution: Choose Relevant, New Item Topics and Attend to Item Difficulty

To address concerns regarding too-easy scales and panel conditioning, we suggest researchers construct novel PK batteries, attending specifically to item difficulty. Additionally, by choosing content that is relevant to the study at hand, political knowledge items will serve as theoretically-motivated moderators in survey experiments, as they more closely assess respondents’ knowledge of the topic at hand. Yet, as demonstrated in Figures 2 and 3, choosing a set of items that is too easy or too hard can be extraordinarily problematic.

So, how can researchers assess likely item difficulty before fielding a scale? As shown in Figure 4, items vary greatly in not only their difficulty, but also in the amount of coverage in the news. Using Google News counts, we estimated the number of news stories likely to contain information about the item’s content published online in the month and year prior to our survey. Unsurprisingly, we find a positive relationship, although with diminishing returns, between the number of Google News stories likely to inform about a particular item and the average score on that item.¹² When building a scale, one may thus exploit publicly-available measures of news coverage to gauge items’ likely difficulty.

In addition to varying item difficulty, one should also aim to choose topics across a variety of typical areas of political knowledge (if a measure of general political awareness is desired). In Study 1, we assessed knowledge across seven different topic areas. These topic areas cover content that one may wish to tailor knowledge items to, depending on the purpose of

¹²In this analysis, we estimate “proportion correct” across true/false and multiple choice formats because we can only use Google News counts to stand in for the item content’s difficulty, not other contributors to item difficulty (e.g., format). We use general Google results for civics items, due to their time-invariant nature.

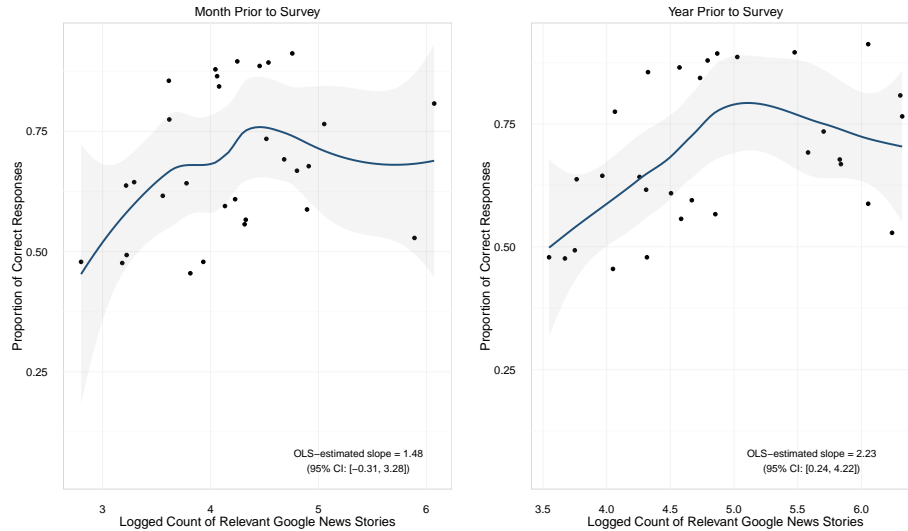


Figure 4: Relationship between political knowledge and Google News [Study 1]
 NOTE: LOESS smoother with 95% confidence intervals depicted in gray.

the study at hand. For example, these seven (six political) topics are of varying relevance for various experiments on policy perceptions, candidate evaluation, party and ideological stereotypes, and general news awareness. If a study wishes to use political knowledge as a moderator in an experiment on judgments of policy, knowledge items that tap domestic and international policy may be of interest. For candidate evaluation, recognition of political persons, domestic affairs, and mass politics may be most relevant. For assessing the moderating role of knowledge on use of party and ideological stereotypes, tapping respondents' knowledge of party positions and broader civics items may be of most importance. Finally, studies relating to general news surveillance may wish to use a mix of items, partially tapping specific media domains, including even entertainment news.

Unsurprisingly, Table 1 demonstrates that all knowledge types considered are positively related, and quite strongly. Also unsurprisingly, the lowest inter-battery correlations exist between the entertainment battery and the other political batteries. This highlights the importance of varying question topics across a variety of content, as the correlations are not perfect. Yet, if one wishes to have a general knowledge scale with relatively few items, the relatively high correlations between topic areas imply that one should prioritize likely item difficulty if facing a tradeoff between coverage and an ideal difficulty distribution.

	Civics	People	Domestic	International	Mass Pols.	Party Positions
People	0.52					
Domestic	0.51	0.44				
International	0.45	0.49	0.45			
Mass Politics	0.41	0.46	0.45	0.49		
Party Positions	0.46	0.42	0.53	0.43	0.43	
Entertainment	0.38	0.41	0.39	0.38	0.48	0.41

Table 1: Correlations between individual knowledge batteries [Study 1]

NOTE: All bivariate correlations statistically significant, $p < .001$. Both true/false and multiple choice questions are included here.

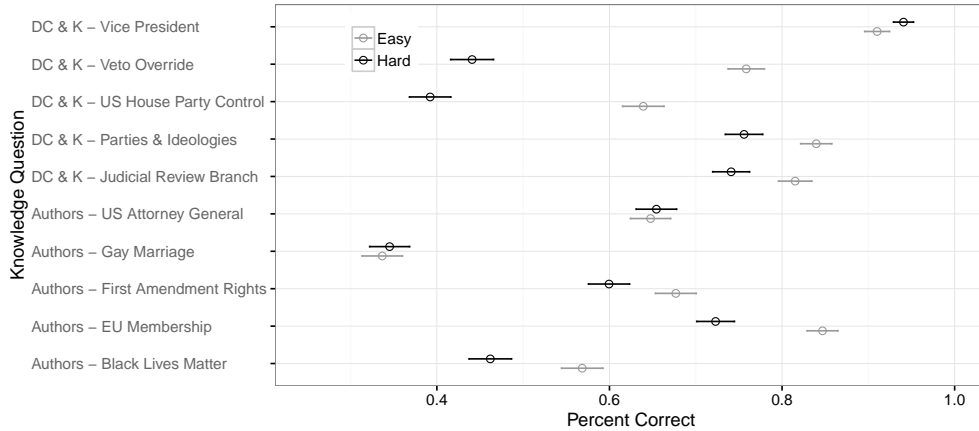


Figure 5: Variation in Difficulty by Response Options [Study 2]

Once the researcher has selected item topics based on Google News counts—or another proxy for likely item difficulty—that are theoretically related to the study at hand and vary maximally in difficulty, one must develop questions from that content. Not only may the particular factual content chosen from news events affect item difficulty, but subtle item design choices may also produce wide variation in difficulty. As shown in Figure 5, even relatively minor changes to the response options provided for a multiple choice item can produce highly variable scores on items. (See OA-2, OA-3, and OA-4 for exact question wordings.) While not divergent for every item, exact question wordings can and do produce divergent scores.

Variation in difficulty arising from question content can be due to more than just varied response options, however. As shown in Figure 6, variation in how a question is asked can produce variability in the difficulty of particular items. Each of the grouped items asked a factually identical question to the others in its group, but with different wording. (E.g., “The

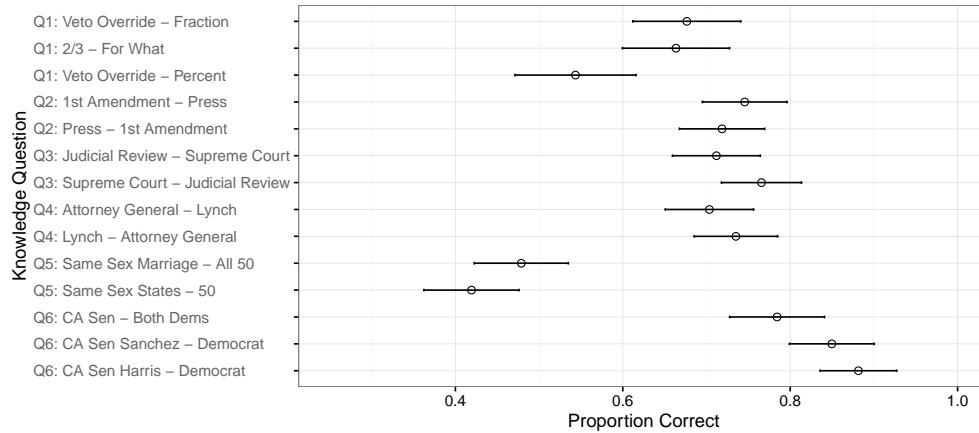


Figure 6: Variation in Difficulty by Question Content [Study 3]

NOTE: Content before the dash is the fact contained in the question, after the dash is the correct answer.

First Amendment guarantees which of these rights?” vs. “Freedom of the press is guaranteed under which amendment?”) Of note, however, when questions were simply statements where the subject and object were swapped from the question to answer, as in questions 2, 3, and 4, variation in difficulty was minimized.

Because of this, we suggest a simple way to generate questions that assess knowledge of a given political topic while minimizing non-topical contributions to item difficulty: generate a simple declarative statement with a subject and object that are central to the particular news story. Because this statement is falsifiable, it is testable as a true/false item, and distractor options can easily be generated by substituting in or out either the subject or object. Furthermore, by keeping the verb constant and testing purely on a relation between a subject and object, ambiguities regarding the degree of truth or characterizations of actions are left off the table. These simple statements help to maximize discrimination of any particular item, as they are simple for respondents to answer and introduce no additional variability in respondents’ ability to accurately record their answer. This approach has additional advantages, including minimizing the researcher degrees of freedom in assessing citizen competence (e.g., Lupia 2016), which we discuss in the following section.

Choosing Item Format

Nearly all political knowledge questions in modern surveys, particularly online surveys, utilize a closed-response format for ease of scoring. Two common approaches are a four-choice multiple choice format and a true/false format. Because respondents are given alternatives, some responses that are scored as correct are almost always due to chance guessing. Dealing with this fact has been the focus of many studies across disparate literatures (e.g., Burton 2001). Chance guessing is not necessarily problematic if we are only concerned about a single question or battery of questions scaled together. However, if we wish to compare different items to each other—whether they share the same number of possible response options or not—comparison becomes nearly impossible without some way of dealing with the systematic error introduced by chance guessing. Because respondents to a four-option question have a 25% chance of guessing correctly (assuming they are guessing at random), while respondents to a separate two-option question have a 50% chance of guessing correctly, we cannot easily compare the results of these two questions to assess difficulty. Additionally, even if we wish to compare two four-item questions, if respondents are not guessing at random, the relative levels of chance guessing make clean comparisons impossible.¹³

¹³Despite extensive study, the discipline lacks a method to head off these concerns. There are two general ways of addressing guessing after data collection. The first, and simplest, is to assume respondents are guessing at random and multiply all scores by $\frac{1-n}{n}$, where n is equal to the number of response options. Thus, a perfect score on a 4-choice question will become 0.75, and one on a 2-choice question will become 0.50. This is problematic for many reasons, as we know guessing is not random: few items exhibit an “as-if random” pattern across the three incorrect answer choices (see OA-2.2). The second general approach is to model the distribution of guessing, which can take several forms. The first is to use general knowledge about response patterns to questions (e.g., an increased propensity to pick a middle option) to specify a vector of probabilities over response options. Alternatively, one can use a three-parameter IRT model to estimate a guessing parameter for each question. However, because

The solution: Use Randomized True/False Items

Our approach to guessing is to reduce its influence at the design stage, rather than after data is collected. The general logic is simple: reduce the degrees of freedom introduced by a researcher into the question format. Because traditional four-option multiple choice questions rely upon the researcher writing the question to come up with three response options besides the correct one, they often unwittingly create questions that can vary dramatically in difficulty due to the increased difficulty of a particular wrong response.¹⁴ By relying upon true/false questions, the original question is transformed into a statement, with either the correct answer in the statement, or one incorrect alternative. Thus, a researcher must only come up with one credible alternative response, rather than three or more.¹⁵

Study 1 provides ample evidence to compare this true/false approach to the more traditional four-choice format.¹⁶ Figure 7 displays respondents' raw accuracy for true/false

these parameters vary question-by-question in response to the actual response distribution, it does not solve the problem of comparability between questions. Figure OA-1 displays both multiple choice and true/false scores with a uniform guessing correction applied and briefly discusses these results.

¹⁴For examples, see the distributions of responses to the “current Attorney General” and “current Secretary of State” multiple choice questions in the Appendix (among others), in which respondents were far more likely to select the penultimate office-holder as an incorrect response than the other two incorrect response.

¹⁵For questions about political party in the US context, this is also useful as there are only two major parties.

¹⁶There are far more robust ways of structuring true/false questions, including presenting both a true and false version of every statement, requiring respondents to pick the correct one. These approaches are sometimes described, in varying styles and permutations, as a multiple true-false item format (Frisbie 1992). This format effectively resembles a two-choice multiple choice format, and allows for a greater role of respondent recognition over recall.

questions, by whether the question as worded was true or false, and Figure 8 displays respondents' raw accuracy for multiple choice.¹⁷ Both figures reveal that our battery of items ranged from rather difficult questions (close to or at the levels we'd expect from guessing) to items that appear far easier, for both formats. Notably, no multiple choice items, even very difficult ones, displayed levels of guessing indistinguishable from those we'd expect from pure guessing. This is likely due to the fact that certain response options provided were easy to rule out, increasing the probability of correct response from random guessing. However, for the true/false versions of the same questions, we find that ten (of the 36 political items) were indistinguishable from pure guessing, and for two, worse than chance. Notably, all but two of these ten were "false" true/false statements. While true/false questions are advantageous on many metrics, this implies an acquiescence bias in response patterns, with respondents more likely to answer "true" than "false" across questions. However, this can be easily remedied by writing two versions of the question—one "true" and one "false"—and randomizing them at the item-respondent level, leading this bias to cancel out in the aggregate.

Additionally, there are other reasons to prefer the true/false format. One obvious metric is survey time, which directly affects the cost of fielding studies online. Figure 10 displays the average time (in seconds) for each of the topical batteries, by question format. True/false questions take respondents significantly less time to answer than multiple-choice questions—in Study 1, a notable 38.3% less time. The average time taken to complete the multiple choice battery was 349 seconds (SE=12.6), compared to just 215 seconds (SE=7.4) for true/false. These differences, whether trimmed for outliers, or with the raw data, as presented, are highly statistically significant, $t(402) = 9.4, p < .001$. These average times are for completing all 42 knowledge questions, averaging to 8.3 seconds per question in the multiple choice format versus 5.1 for true/false. Thus, a researcher can ask far more questions, covering more topics,

However, we present the simplest, easiest to implement question format in the current study.

¹⁷These are plotted separately because of the difficulty in comparing the two quantities because of differences due to guessing.

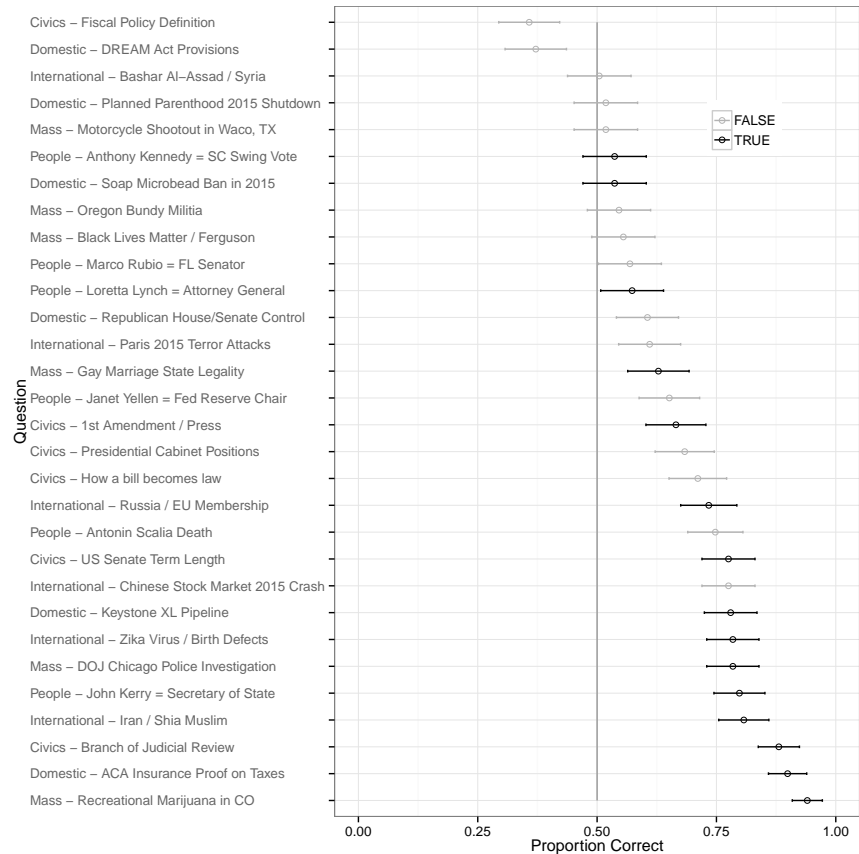


Figure 7: Raw Question Accuracy: True/False Questions [Study 1]

NOTE: 95% Confidence Intervals shown. The vertical gray line indicates the level of accuracy expected by guessing. Questions are colored by whether the provided statement was true or false. See the Appendix for exact question wording. Entertainment questions not reported here, as they were not scaled into the political knowledge full scale. Party placement questions also not reported, as they are a two-option multiple choice.

in the same amount of survey space.

Of course, true/false items do result in slightly lower inter-item reliability, 0.77 versus 0.86 in the full 36-item battery.¹⁸ However, scale length appears to affect reliability far more

¹⁸Given each item has a mean closer to 0.5 in each scale, indicating greater variance in the response distribution, this result is unsurprising. Using the three scales in Figure 2, we find that the optimal battery has a Cronbach's $\alpha = 0.35$ while the too-hard battery $\alpha = 0.47$ and the too-easy battery $\alpha = 0.66$. Thus, a more optimally spread short battery will inevitably decrease inter-item reliability. A better test, only possible with panel data, would be calculating the test-retest reliability. However, even that is problematic due to changes in the information environment and possible panel conditioning over multiple waves.

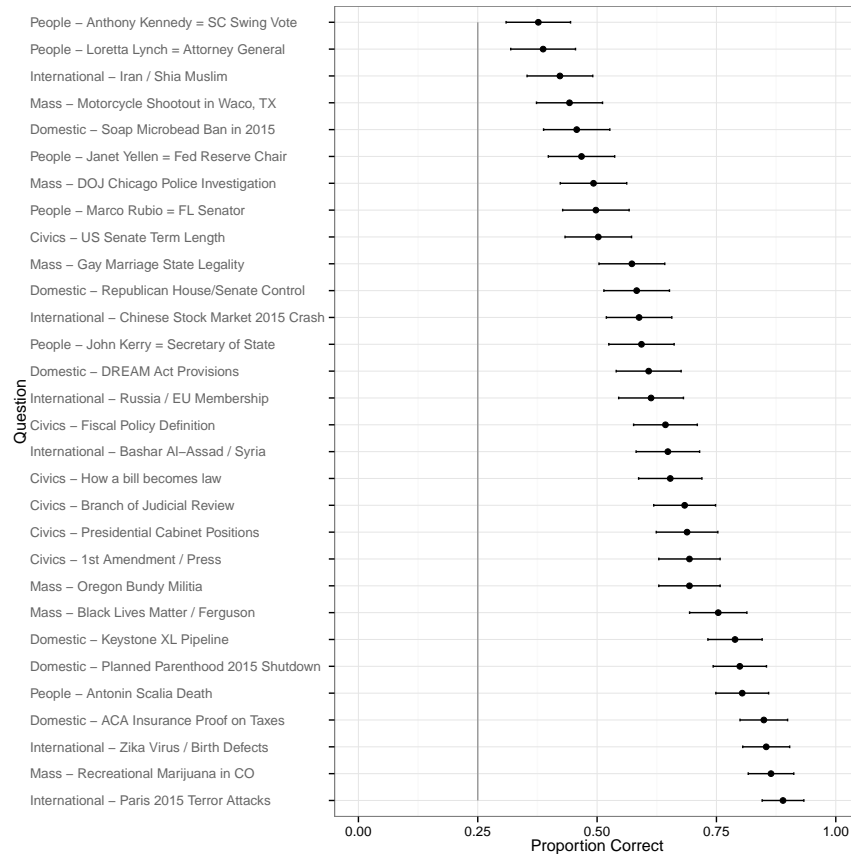


Figure 8: Raw Question Accuracy: Multiple Choice Questions [Study 1]

NOTE: 95% Confidence Intervals shown. The vertical gray line indicates the level of accuracy expected by guessing. See the Appendix for exact question wording. Entertainment questions not reported here, as they were not scaled into the political knowledge full scale. Party placement questions also not reported, as they are a two-option multiple choice.

than item format. To better understand characteristics of these scales as they vary in length and item format, we conducted a series of simulations using the Study 1 data. From the full set of 36 political knowledge items, we randomly chose a subset—either 3, 5, 10, 15, 20, 25, or 30 items—and then calculated a number of statistics about the performance of that scale.¹⁹ Our simulations also allow us to observe how scale properties change as we move from extremely short scales (3-item) to the full 36-item battery. Figure 9 displays average score and scale reliability by item format scale length. (See Figure OA-4 for plots showing IRT difficulty and discrimination parameters as well.) While average score does not change in expectation, the number of items in a battery quickly increases the precision with

¹⁹This was repeated 1,000 times to create distributions of possible scales of each length.

which we estimate a respondent’s political knowledge, for both item formats. With respect to reliability, we see obvious increases as we have longer scales, with only minor reliability gains for the multiple choice format, *ceteris paribus*.

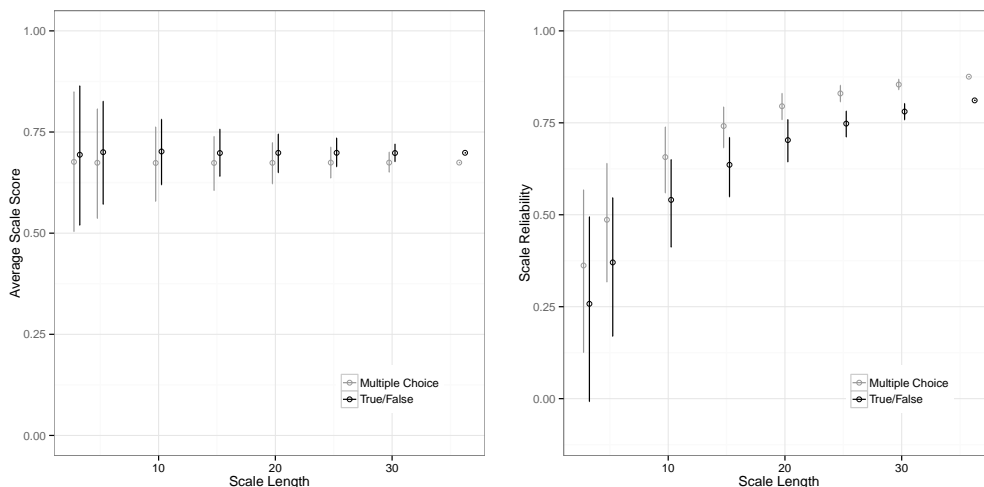


Figure 9: Scale Length, Question Type, and Battery Properties [Study 1]

NOTE: 95% Confidence Intervals are bootstrapped from a sample of 1000 draws of that scale length from our original 36 political items.

As in the full battery, the reliabilities for each individual topic battery are greater under the multiple choice format. Because of their relatively short (six-item) scale length, these differences appear quite large, as implied by Figure 11.²⁰ However, in terms of aggregate difficulty, there are few marked differences in average scores by true/false or multiple choice format, as shown in Figure 12. Furthermore, because of the efficiency gains from the true/false format, the difference in reliability between an m -item true/false battery and a n -item multiple-choice battery, holding survey time/space allocated to the PK battery constant, is smaller than the differences observed in Figures 9 and 11, where $m = n$.

²⁰This difference is statistically significant, $p < .01$. Note that, in this case, Cronbach’s alpha is both a function of the increased likely guessing on true/false questions, as well as simply the number of response options and the marginal distribution of scores it produces.

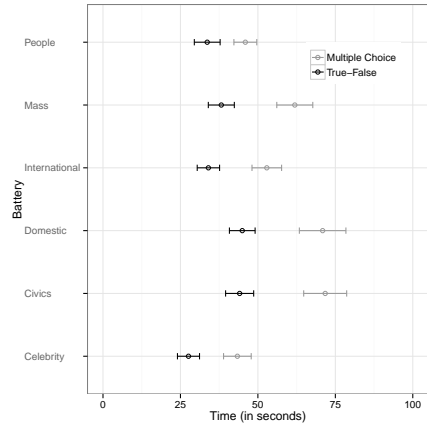


Figure 10: Question Battery: Time by True/False or Multiple Choice [Study 1]
NOTE: 95% Confidence Intervals shown.

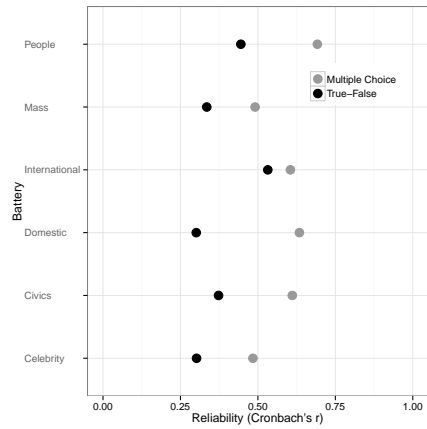


Figure 11: Question Battery: Reliability by True/False or Multiple Choice [Study 1]

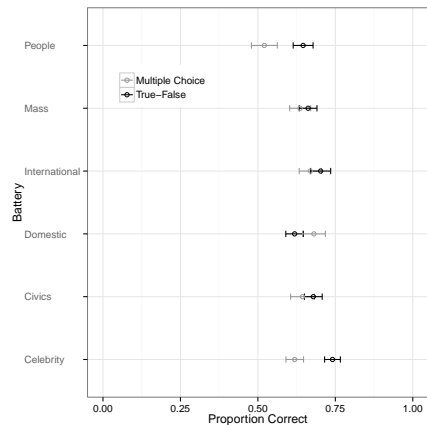


Figure 12: Question Battery: Percent Correct by True/False or Multiple Choice, Uncorrected for Guessing [Study 1]

NOTE: 95% Confidence Intervals shown. Party Placement is not shown because it does not fall neatly into either question type.

Scaling Items Together

Once a scale is chosen and administered, how should one score respondents? The use of batteries, as opposed to single items, is obviously paramount; as classical test theory holds, responses to an individual recall item are a function of the respondent’s ability, the item’s difficulty and discriminatory power, and random error. As we average across multiple items, random measurement error cancels out and we may thus more precisely estimate the respondent’s ability (e.g., Ansolabehere, Rodden and Snyder 2008).

To evaluate how well various knowledge items scale together, Table 2 shows the inter-item relationships for both our five-item scale and the classic Delli Carpini & Keeter five-item scale from Study 2. Notably, while distributionally different (with our scale having higher variance and a lower average score), the items appear to scale together quite similarly. That is, when items are ordered by difficulty, the resultant ordering is preserved when considering *patterns* of correct responses. This is most apparent by examining the rows from top to bottom and columns from left to right.

		Authors				
	Russia/EU	Lynch/AG	1stAmnd/Press	BLM/MO	GayMarriage/50	
Russia/EU	0.784					
Lynch/AG	0.545	0.651				
1stAmnd/Press	0.518	0.435	0.637			
BLM/MO	0.42	0.367	0.342	0.516		
GayMarriage/50	0.283	0.243	0.251	0.195	0.341	

		Delli Carpini & Keeter				
	VP/Biden	Dem=Lib/Rep=Cons	SC/JudicialRev	VetoOverride/2/3	HouseControl/Rep	
VP/Biden	0.926					
Dem=Lib/Rep=Cons	0.755	0.799				
SC/JudicialRev	0.737	0.635	0.777			
VetoOverride/2/3	0.573	0.507	0.479	0.602		
HouseControl/Rep	0.492	0.442	0.428	0.341	0.513	

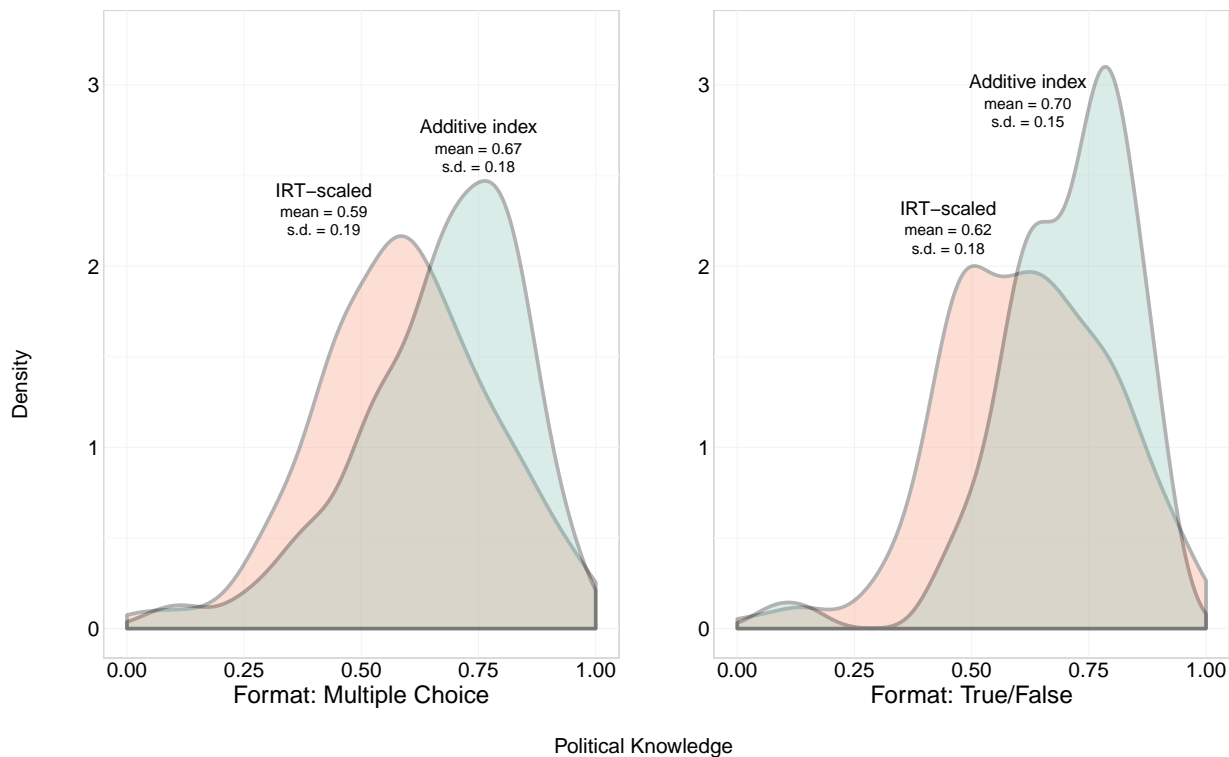
Table 2: Inter-Item Correctness [Study 2]

NOTE: Bold proportions indicate proportion correct on that item. All other proportions represent the proportion of respondents getting both items correct.

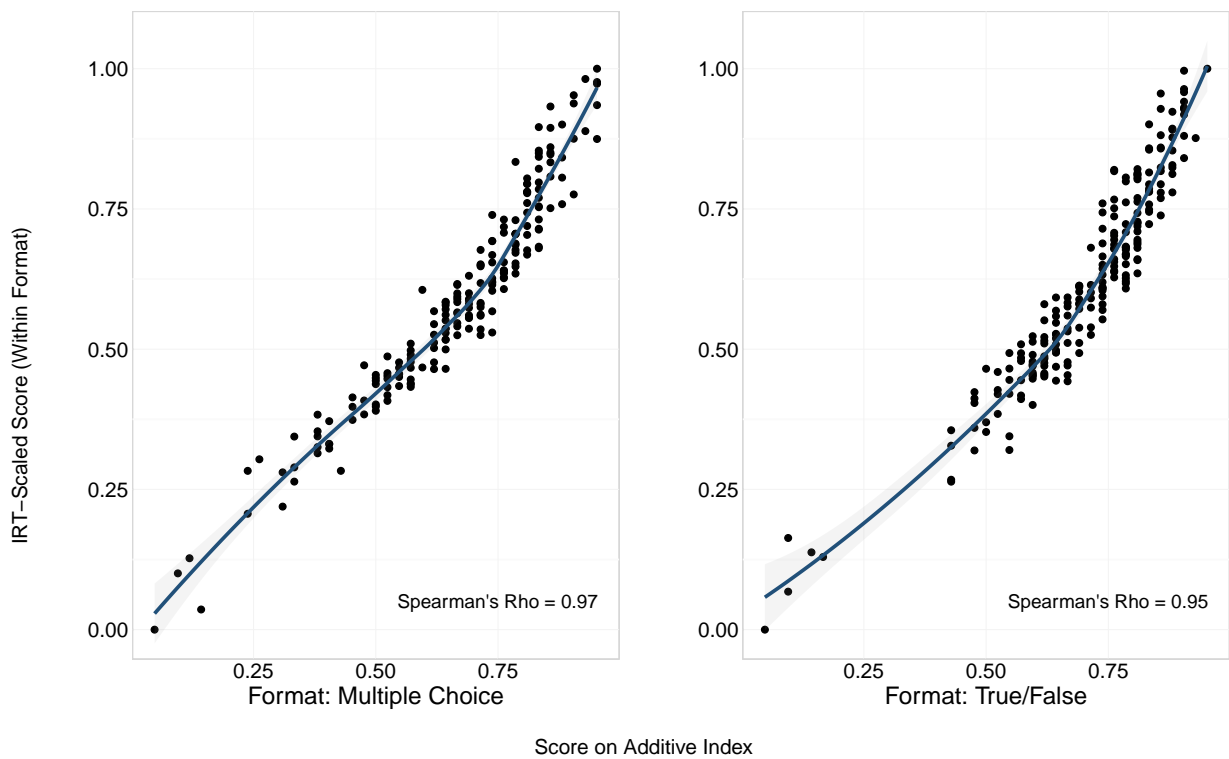
While the simplest way is adding or averaging, does the use of IRT provide better performance? Put simply, if the scale is well-designed to begin with, no. Figure 13a implies that IRT-scaling does transform the distribution of scores in ways that give it more ideal properties, *on average*; in combining all 36 items into one political knowledge scale, IRT-

Figure 13: Scaling Political Knowledge [Study 1]

(a) Distribution of Political Knowledge, by Question Format and Calculation Approach



(b) Comparing Scores on Additive and IRT-Scaled Knowledge Indices



scaling improved spread and diminished right-skew. But if one chooses items thoughtfully, with those goals in mind, the marginal returns from IRT-scaling will be more modest. Further, Table OA-3 shows both bivariate correlations between common political variables and our scale, both additive and IRT, broken down by question type, as well as multiple regressions predicting the scores. We see few differences in the observed relationships regardless of scaling procedure, for both multiple choice and true/false batteries. In sum, IRT-scaling may help to rescale poorly-constructed knowledge batteries, but in an assumption-laden way (Broockman 2016; Lupia 2016); thoughtful battery construction at the outset will yield similar desirable properties without the need for a latent variable model.

Despite the differences in time and reliability, true/false and multiple choice items appear to produce similar distributions of political knowledge, as shown in Figure 13. The kernel density estimates plotted in cyan represent scores as typically constructed from individual items, that is, via a simple, unweighted additive scale. As the plots show, additive scales based on true/false items have a slightly more compact distribution, given the increased probability of getting an item correct due to guessing. However, when examining distributions of IRT-scaled scores (in salmon), the two types of items yield highly similar distributions. On the one hand, this implies that batteries based on true/false items benefit more from IRT scaling. Again, though, Figure 13a presents the distribution of scores across all items administered, rather than optimal batteries. Furthermore, with respect to multiple choice and true/false items' performance in IRT-scaling, we find that the two formats yield similar difficulty and discrimination parameters in IRT models.²¹

In a very reassuring finding for those using more common self-reported measures of political interest, the best predictor of all knowledge scores in Study 1 is self-reported political knowledge. Interestingly, we see no apparent effect of self-reported time spent on Mechanical

²¹These are shown in Figures OA-2 and OA-3 in the Appendix, respectively. The Spearman's rho rank-order correlation coefficient for the difficulty parameters is 0.57. It is somewhat lower, although still moderately strong 0.42 for the discrimination parameters.

Turk per day, suggesting less difference between more professionalized Mechanical Turk workers and more casual users than selection and panel conditioning hypotheses suggest. However, Mechanical Turk workers clearly have varied interest in politics, and this shows. At the completion of the survey, we asked respondents if they wished to be given the answers to the questions, as well as their own scores, once the survey had closed for all respondents. While not reported in the table, this item is highly predictive of performance: respondents who wanted their score and the answers emailed to them scored 7.1 percentage points higher on average, 73.9% vs. 66.9%, $t(396) = 4.8, p < .001$.

Discussion

Political behavior research frequently relies on the concept of political knowledge, and rightly so: it is fundamental to citizens' holding elected officials accountable. Nevertheless, it is often measured haphazardly or with batteries optimized decades ago. In this paper, we have demonstrated the pitfalls of this approach. Not only can it undermine validity of measurement (e.g., Barabas et al. 2014), but it can also obscure meaningful relationships. With Monte Carlo simulations, we demonstrated that too-easy and too-hard batteries introduce attenuation bias into the well-documented relationship between political knowledge and attitudinal constraint, and in predictable ways. With a randomized controlled experiment, we demonstrated that a carefully constructed, contemporary battery similarly outperforms the canonical (and still frequently used) battery of the "telephone era" in survey research (Delli Carpini and Keeter 1996). While one may be tempted to retain existing batteries to bridge across time, the immense variation in item format and response options already makes these comparisons problematic. As such, our primary recommendation is as follows: rather than constructing batteries off the top of one's head or pulling existing batteries off the shelf, researchers should design their own batteries, of optimal difficulty for their sample and with items that tap the specific domain(s) of political knowledge their theory engages. Many of

these tailored batteries might incorporate “existing” items in our specified format—we do not inherently fault any existing items, purely the outcomes that result when adopting them in a haphazard fashion and format.

This may appear easier said than done. We identified our “carefully constructed” batteries from a 36-item battery, simulating 5- and 6-item batteries from those 36 items, and optimizing on battery properties (average item difficulty and spread of scores). But items are unlikely to maintain their “2016 difficulty and discriminatinon” over time—the political agenda and news cycle shift constantly—so how can researchers construct optimal batteries without replicating this paper’s exercise every time they measure political knowledge? As we show using Google News counts, the media coverage an item’s topic receives predicts the item’s difficulty well (with diminishing marginal returns). Thus, researchers may use media metrics to construct political knowledge batteries with items of varying difficulty. Not only does building unique batteries this way improve validity and performance of measures, but it also attenuates concerns over panel conditioning among online respondents who take many political surveys.

We further recommend the use of the true/false format over the multiple choice format. Although multiple choice yields slight reliability gains—holding the number of items administered constant—the true/false format allows researchers to administer more items in a given amount of time. Furthermore, it reduces the variability of difficulty in items that can be constructed from any particular news story or political fact. This latter advantage reduces “researcher degrees of freedom” in the measurement of political awareness and allows for increased comparability between items on their contents’ actual difficulty. Aside from these differences, the two formats perform similarly on other desirable metrics.

Measurement is paramount in the social sciences. It is also difficult. On the one hand, measurement must be systematic. To discover meaningful relationships, and for collaboration and debate to bear fruit, we require rigorous methods and agreement on standards. On the other hand, it has been said that we study “clouds, not clocks” (Almond and Genco 1977;

Popper et al. 1972)—that is, we can make sense of the social world, but its empirical realities are more complex and transitory than those that govern the physical universe. As such, the social sciences require greater flexibility in measurement than the natural sciences. This flexibility need not be at odds with scientific standardization. As we have shown, rigorous but flexible *procedures* can protect against false negatives from rigid *instrumentation*, while also ensuring integrity of measurement.

References

- Adcock, Robert and David Collier. 2001. "Measurement Validity: A Shared Standard for Qualitative and Quantitative Research." *American Political Science Review* 95(3):529–546.
- Ahler, Douglas J., Jack Citrin, Michael C. Dougal and Gabriel S. Lenz. 2017. "Face Value? Experimental Tests of the Influence of Candidate Appearance on Electoral Choice." *Political Behavior* 39(1):77–102.
- Aldrich, John H., John L. Sullivan and Eugene Borgida. 1989. "Foreign Affairs and Issue Voting: Do Presidential Candidates "Waltz Before A Blind Audience?"." *American Political Science Review* 83(1):123–141.
- Almond, Gabriel A and Stephen J Genco. 1977. "Clouds, clocks, and the study of politics." *World politics* 29(04):489–522.
- Althaus, Scott L. 1998. "Information Effects in Collective Preferences." *American Political Science Review* 92(3):545–558.
- Ansolabehere, Stephen, Jonathan Rodden and James M. Snyder. 2008. "The Strength of Issues: Using Multiple Measures to Gauge Preference Stability, Ideological Constraining, and Issue Voting." *American Political Science Review* 102(215-232).
- Barabas, Jason, Jennifer Jerit, William Pollock and Carlisle Rainey. 2014. "The Question(s) of Political Knowledge." *American Political Science Review* 108(4):840–855.
- Bartels, Larry M. 1996. "Uninformed Votes: Information Effects in Presidential Elections." *American Journal of Political Science* 40(1):194–230.
- Bartle, John. 2000. "Political Awareness, Opinion Constraint and the Stability of Ideological Positions." *Political Studies* 48(3):467–484.
- Baum, Matthew A. 2003. "Soft News and Political Knowledge: Evidence of Absence or Absence of Evidence." *Political Communication* 20(2):173–190.

- Berinsky, Adam J, Gregory A Huber and Gabriel S Lenz. 2012. "Evaluating online labor markets for experimental research: Amazon. com's Mechanical Turk." *Political Analysis* 20(3):351–368.
- Brewer, Paul R. 2003. "Values, Political Knowledge, and Public Opinion about Gay Rights: A Framing-Based Account." *Public Opinion Quarterly* 67(2):173–201.
- Broockman, David E. 2016. "Approaches to Studying Representation." *Legislative Studies Quarterly* Forthcoming.
- Burton, Richard F. 2001. "Quantifying the effects of chance in multiple choice and true/false tests: question selection and guessing of answers." *Assessment & Evaluation in Higher Education* 26(1):41–50.
- Clifford, Scott and Jennifer Jerit. 2016. "Cheating on Political Knowledge Questions in Online Surveys An Assessment of the Problem and Solutions." *Public Opinion Quarterly* p. nfw030.
- Converse, Philip. 1964. The Nature of Belief Systems in Mass Publics. In *Ideology and Discontent*, ed. David Apter. New York: Free Press.
- Converse, Philip E. 1975. Public opinion and voting behavior. In *Handbook of political science: Nongovernmental politics*, ed. Nelson W Polsby and Fred I Greenstein. Vol. 4 Addison-Wesley Reading, Mass. pp. 75–169.
- Cronbach, Lee J. and Paul E. Meehl. 1955. "Construct Validity in Psychological Tests." *Psychological Bulletin* 52(4):281–302.
- Curtin, Richard, Stanley Presser and Eleanor Singer. 2005. "Changes in Telephone Survey Nonresponse Over the Past Quarter Century." *Public Opinion Quarterly* 69(1):87–98.
- Delli Carpini, Michael X. and Scott Keeter. 1996. *What Americans Know About Politics and Why It Matters*. New York: Yale University Press.

- Frisbie, David A. 1992. "The Multiple True-False Item Format: A Status Review." *Educational Measurement: Issues and Practice* 11(4):21–26.
- Galston, William A. 2001. "Political Knowledge, Political Engagement, and Civic Education." *Annual Review of Political Science* 4:217–234.
- Hillygus, D. Sunshine, Natalie Jackson and McKenzie Young. 2014. Professional Respondents in Non-Probability Online Panels. In *Online Panel Research: A Data Quality Perspective*, ed. Mario Callegaro, Reg Baker, Jelke Bethlehem, Anja S. Goritz and Jon A. Krosnick. New York: John Wiley & Sons pp. 219–237.
- Hoyle, Rick H., Monica J. Harris and Charles M. Judd. 2002. *Research Methods in Social Relations*. 7th ed. Toronto: Wadsworth Thomson Learning.
- Johnson, Paul. 2009. What Knowledge is of Most Worth? In *The Political Psychology of Democratic Citizenship*, ed. Eugene Borgida, Christopher M. Federico and John L. Sullivan. New York: Oxford University Press.
- Krupnikov, Yanna and Adam Seth Levine. 2014. "Cross-Sample Comparisons and External Validity." *Journal of Experimental Political Science* 1:59–80.
- Lenz, Gabriel S. 2012. *Follow the Leader? How Voters Respond to Politicians' Policies and Performance*. Chicago: University of Chicago Press.
- Lenz, Gabriel S. and Chappell Lawson. 2011. "Looking the Part: Television Leads Less Informed Citizens to Vote Based on Candidates' Appearance." *American Journal of Political Science* 55(3):574–589.
- Lenz, Gabriel S., Shad Turney and Sean Freeder. 2016. "The Importance of Knowing "What Goes With What"—Reexamining the Evidence on Attitude Stability, Policy Voting, and Multi-Item Issue Scales." Working Paper, University of California, Berkeley.
- Lippman, Walter. 1922. *Public Opinion*. Transaction Publishers.

- Lupia, Arthur. 2006. "How Elitism Undermines the Study of Voter Competence." *Critical Review* 18(1-3):217–232.
- Lupia, Arthur. 2016. *Uninformed: Why People Know So Little About Politics and What We Can Do About It*. New York: Oxford University Press.
- Luskin, Robert C. 1987. "Measuring Political Sophistication." *American Journal of Political Science* 31(4):856–899.
- Miller, Warren E. and Donald E. Stokes. 1963. "Constituency Influence in Congress." *American Political Science Review* 57(1):45–56.
- Mondak, Jeffrey J. 2001. "Developing Valid Knowledge Scales." *American Journal of Political Science* 45(1):224–238.
- Nunnally, Jim C and IH Bernstein. 1994. *Psychometric theory*. New York: McGraw-Hill.
- Nyhan, Brendan and Jason Reifler. 2010. "When Corrections Fail: The Persistence of Political Misperceptions." *Political Behavior* 32(2):303–330.
- Popper, Karl Raimund et al. 1972. "Objective knowledge: An evolutionary approach."
- Roush, Carolyn E. 2016. Believing the Worst: Out-Party Hostility and Receptiveness to Political Misinformation. In *Annual Meeting of the American Political Science Association*. Philadelphia: .
- Schudson, Michael. 1998. *The Good Citizen: A History of American Civic Life*. New York: Free Press.
- Thompson, Stephen K. 2012. *Sampling*. 3rd ed. New York: Wiley.
- Zaller, John R. 1992. *The Nature and Origins of Mass Opinion*. Cambridge: Cambridge University Press.

Online Appendix

OA-1 Supplementary Tables/Figures

Question	% Respondents				
Gender	52.4%	42.9%			
	Male	Female			
Race/Ethnicity	73.0%	9.7%	7.1%	3.6%	1.7%
	White/Caucasian	Black/Af Am	Asian/PI	Hispanic/Latino	Other
Education	1.4%	25.1%	18.7%	38.6%	11.4%
	Less than HS	High School	Some College	College Deg.	Advanced Deg.
Civics Course	33.6%	39.6%	21.6%		
	More than One	One	None		
Political Interest	33.9%	45.7%	14.0%	5.7%	
	Most of the time	Some of the time	Only now and then	Hardly at all	
News Sources	30.3%	24.2%	44.5%	46.0%	39.8%
	Local Newspaper	Nat'l Newspaper	Broadcast TV	Cable TV	Local TV
	63.7%	79.4%	46.0%		
	Social Media	Internet	Family/Friends		
Pocketbook (Past Year)	5.7%	27.5%	41.0%	13.5%	6.6%
	Much Better	Somewhat Better	About the Same	Somewhat Worse	Much Worse
Sociotropic (Past Year)	3.1%	27.0%	42.9%	17.3%	4.3%
	Much Better	Somewhat Better	About the Same	Somewhat Worse	Much Worse
Primary Elec Turnout	32.2%	15.4%	14.0%	9.0%	24.2%
	Extremely Likely	Very Likely	Moderately Likely	Slightly Likely	Not Likely
General Elec Turnout	49.1%	19.2%	10.2%	5.9%	10.4%
	Extremely Likely	Very Likely	Moderately Likely	Slightly Likely	Not Likely
2016 Vote Choice	5.0%	5.5%	5.5%	13.7%	2.6%
	Kasich	Rubio	Cruz	Trump	Carson
	14.2%	38.9%	9.2%		
	Clinton	Sanders	None of Above		
Party ID (3-point)	50.5%	16.1%	28.7%		
	Democrat	Independent	Republican		
Party ID (7-point)	17.1%	23.0%	10.4%	16.1%	
	Strong Democrat	Not Strong Dem.	Lean Democrat	Independent	
	8.1%	14.7%	5.9%		
	Strong Republican	Not Strong Rep.	Lean Republican		
Ideology	7.8%	25.6%	12.6%	23.0%	
	Extr. Liberal	Liberal	Slightly Liberal	Moderate	
	3.6%	10.2%	12.6%		
	Extr. Conservative	Conservative	Slightly Conservative		
Mturk Usage (Per Day)	19.0%	29.9%	17.3%	11.1%	17.5%
	Less than 1hr	1-2hr	2-3hr	3-4hr	4+hr
Want Score/Answers	40.0%	54.3%			
	Yes	No			

Table OA-1: Mechanical Turk Survey Sample Characteristics, N=422 [Study 1]

NOTE: Questions do not sum to 100% because of rounding and nonresponse. Respondents could check multiple news sources.

Table OA-2: Predictors of Specific Types of Knowledge [Study 1]
(a) Multiple Choice Questions

	Multiple Choice, Additive Score on Individual Batteries:						
	Civics (1)	People (2)	Domestic (3)	International (4)	Mass Pols. (5)	Party Positions (6)	Entertainment (7)
Political correlates							
Democrat	-.02 (.05)	-.00 (.06)	.04 (.05)	.10* (.05)	.11** (.05)	.03 (.04)	.02 (.04)
Republican	.00 (.05)	.03 (.06)	.07 (.05)	.09 (.05)	.09* (.05)	-.01 (.04)	-.05 (.04)
Strong Partisan	-.08* (.04)	-.08* (.05)	-.07* (.04)	-.09** (.04)	-.04 (.04)	-.01 (.03)	-.06* (.03)
Consistency	.16* (.08)	.14 (.09)	.04 (.08)	-.03 (.04)	.06 (.07)	.05 (.06)	.03 (.06)
Turnout Likelihood	.17** (.08)	.07 (.08)	.11 (.07)	-.07 (.07)	.01 (.08)	.00 (.05)	.09 (.06)
Education indicators							
H.S. Degree	.14 (.13)	.16 (.14)	.29** (.12)	.08 (.12)	.03 (.11)	.29*** (.09)	-.05 (.09)
Some College	.17 (.13)	.19 (.14)	.27** (.12)	.10 (.12)	-.00 (.12)	.30*** (.09)	-.06 (.10)
Bachelor's Degree	.18 (.13)	.18 (.13)	.35*** (.12)	.11 (.12)	-.01 (.11)	.35*** (.09)	-.08 (.09)
Advanced Degree	.23* (.14)	.27* (.15)	.34 (.13)	.12 (.13)	-.01 (.12)	.36*** (.10)	-.07 (.10)
Self-reports							
Attn. to campaigns/elections	.08 (.10)	.21* (.11)	-.02 (.09)	.04 (.10)	.20** (.09)	.02 (.07)	.04 (.08)
Attn. to current events	-.03 (.08)	.04 (.08)	.14** (.07)	.02 (.07)	.07 (.06)	.05 (.05)	-.00 (.06)
Attn. to domestic policy	.16 (.11)	.26** (.12)	.24** (.10)	.12 (.10)	.00 (.09)	.05 (.08)	.05 (.08)
Attn. to foreign affairs	-.07 (.09)	-.01 (.09)	-.08 (.08)	.21** (.08)	.00 (.08)	.06 (.06)	-.04 (.06)
Attn. to entertainment news	-.08 (.06)	-.10 (.06)	-.16*** (.06)	-.06 (.06)	.03 (.05)	-.06 (.04)	.09** (.04)
Time on Turk	.04 (.08)	.05 (.09)	.06 (.07)	-.03 (.08)	.01 (.07)	.07 (.06)	-.10 (.06)
Constant	.30 (.14)	-.03 (.15)	.14 (.14)	.38 (.13)	.35 (.13)	.44 (.10)	.63 (.11)
R2	.18	.29	.22	.18	.14	.18	.11
n	189	189	189	189	189	189	189

(b) True-False Questions

	True/False, Additive Score on Individual Batteries:						
	Civics (1)	People (2)	Domestic (3)	International (4)	Mass Pols. (5)	Party Positions (6)	Entertainment (7)
Political correlates							
Democrat	-.06 (.04)	.03 (.04)	.03 (.04)	-.07 (.05)	-.06 (.04)	-.05 (.03)	-.01 (.04)
Republican	-.08* (.05)	-.04 (.05)	.01 (.05)	-.12 (.05)	-.04 (.05)	-.11*** (.04)	.02 (.04)
Strong Partisan	.00 (.03)	.01 (.04)	-.03 (.03)	.00 (.04)	-.04 (.03)	-.02 (.03)	.01 (.03)
Consistency	.16 (.07)	.18** (.07)	.09 (.07)	.02 (.07)	.08 (.07)	.09* (.05)	-.11* (.07)
Turnout Likelihood	-.02 (.07)	-.06 (.07)	.04 (.07)	-.04 (.07)	.01 (.07)	.09 (.05)	-.02 (.06)
Education indicators							
H.S. Degree	-.19 (.14)	-.13 (.15)	.06 (.14)	.06 (.16)	-.06 (.14)	.05 (.11)	.05 (.13)
Some College	-.19 (.14)	-.11 (.15)	.06 (.14)	.03 (.16)	-.03 (.14)	-.03 (.12)	.03 (.13)
Bachelor's Degree	-.13 (.14)	-.12 (.15)	.07 (.14)	.13 (.15)	-.02 (.14)	.07 (.11)	.09 (.13)
Advanced Degree	-.04 (.14)	-.02 (.15)	.05 (.15)	.11 (.16)	.03 (.15)	.06 (.12)	.07 (.13)
Self-reports							
Attn. to campaigns/elections	.02 (.07)	.16** (.08)	.00 (.08)	-.06 (.08)	.21*** (.08)	.05 (.06)	.09 (.07)
Attn. to current events	.01 (.07)	-.13* (.07)	.10 (.07)	.03 (.08)	-.07 (.07)	.02 (.06)	.03 (.07)
Attn. to domestic policy	.16* (.08)	.28*** (.09)	.20** (.09)	.19** (.09)	.16* (.09)	.14* (.07)	.10 (.06)
Attn. to foreign affairs	-.04 (.07)	.00 (.07)	-.13* (.07)	.15** (.07)	-.11 (.07)	-.03 (.05)	-.00 (.08)
Attn. to entertainment news	-.13*** (.04)	-.08* (.05)	-.10** (.04)	-.14*** (.05)	-.14*** (.04)	-.11*** (.04)	.05 (.04)
Time on Turk	-.06 (.07)	-.04 (.07)	.03 (.07)	-.12 (.07)	.08 (.07)	-.03 (.05)	-.00 (.06)
Constant	.82 (.15)	.57 (.16)	.41 (.15)	.61 (.17)	.60 (.15)	.78 (.12)	.56 (.14)
R2	.22	.27	.14	.26	.20	.26	.13
n	210	210	210	210	210	210	210

Table OA-3: Political Knowledge and its Predictors on MTurk [Study 1]**(a) Bivariate Correlations**

	Multiple Choice		True/False	
	Additive Score	IRT-Scaled Score	Additive Score	IRT-Scaled Score
Democrat	.08	.04	.07	.11
Republican	.12	.14	.03	-.02
Strong Partisan	.04	.04	.11	.17
Consistency	.26	.28	.29	.34
Turnout Likelihood	.32	.31	.23	.29
Education	.19	.24	.24	.25
Self-Reported Political Attn.	.46	.50	.44	.47
Self-Reported Entertain. Attn.	-.03	-.06	-.21	-.23
Self-Reported Time on Turk	-.04	-.05	-.08	-.10

(b) Multivariate Regression

	Multiple Choice		True/False	
	Additive Score	IRT-Scaled Score	Additive Score	IRT-Scaled Score
	(1)	(2)	(3)	(4)
Political correlates				
Democrat	.04 (.03)	.04 (.03)	-.02 (.03)	-.03 (.03)
Republican	.03 (.03)	.05 (.03)	-.04 (.03)	-.08** (.03)
Strong Partisan	-.06** (.03)	-.07*** (.03)	-.00 (.02)	.00 (.02)
Consistency	.07 (.05)	.10** (.05)	.07* (.04)	.14*** (.05)
Turnout Likelihood	.07 (.04)	.05 (.05)	.01 (.04)	.04 (.05)
Education indicators				
H.S. Degree	.13* (.07)	.16** (.08)	-.01 (.09)	-.02 (.10)
Some College	.13* (.07)	.17** (.08)	-.02 (.09)	-.04 (.10)
Bachelor's Degree	.14** (.07)	.19** (.08)	.03 (.09)	.01 (.10)
Advanced Degree	.17** (.08)	.23*** (.08)	.05 (.09)	.05 (.10)
Self-reports				
Self-Reported Political Attn.	.26*** (.05)	.32*** (.06)	.23*** (.04)	.27*** (.05)
Self-Reported Entertain. Attn.	-.05 (.03)	-.08** (.04)	-.09*** (.02)	-.12 (.03)
Self-Reported Time on Turk	.02 (.04)	.03 (.04)	-.04 (.04)	-.06 (.05)
Constant	.32 (.05)	.16 (.08)	.60 (.09)	.49 (.11)
R2	.27	.32	.30	.37
n	189	189	210	210

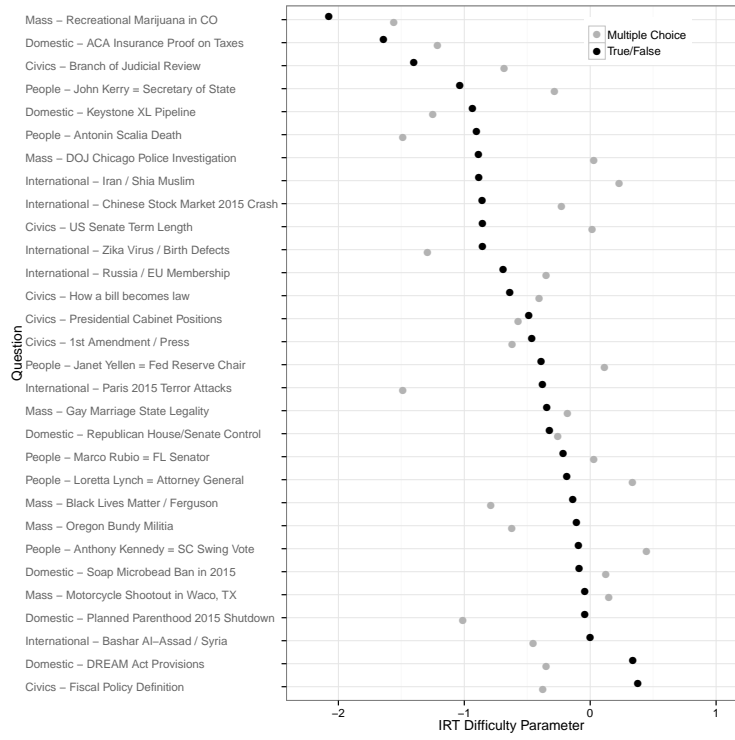


Figure OA-2: Questions: Multiple Choice versus True/False IRT Difficulty Parameters [Study 1]
 NOTE: Spearman's rho = 0.57. See the Appendix for exact question wording. Confidence Intervals based on naive standard errors of the IRT parameters are so small they plot behind the points themselves.

Note that in both Figures OA-2 and OA-3, the plots are sorted based on the true/false parameter, giving somewhat of a false sense that the true/false parameters are more stable. Rather, we aim to show that in nearly all cases, the items have very similar parameters. While a scatterplot could also show this, we prefer this presentation so that one can refer to particular questions

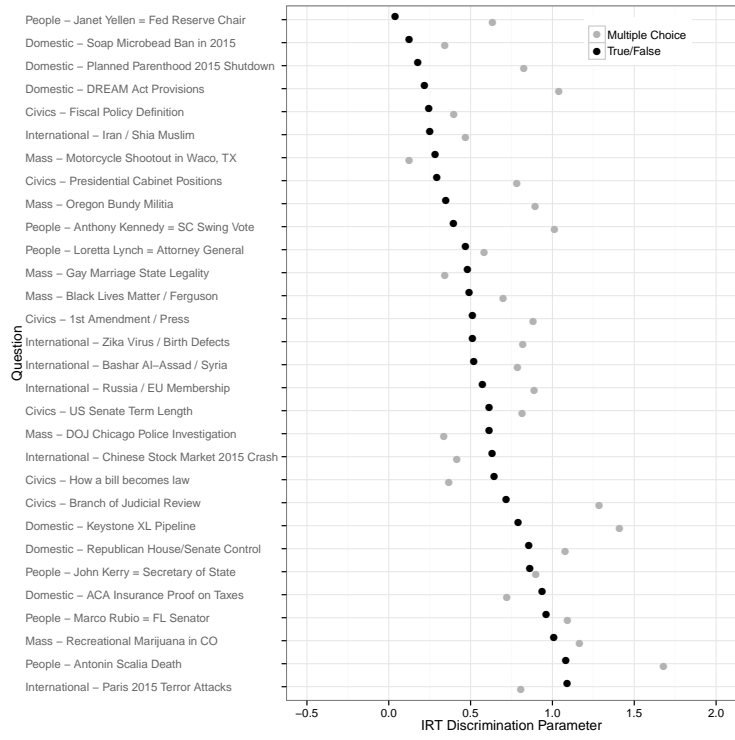


Figure OA-3: Questions: Multiple Choice versus True/False IRT Discrimination Parameters [Study 1]

NOTE: Spearman’s rho = 0.42. See the Appendix for exact question wording. Confidence Intervals based on naive standard errors of the IRT parameters are so small they plot behind the points themselves.

Note that in both Figures OA-2 and OA-3, the plots are sorted based on the true/false parameter, giving somewhat of a false sense that the true/false parameters are more stable. Rather, we aim to show that in nearly all cases, the items have very similar parameters. While a scatterplot could also show this, we prefer this presentation so that one can refer to particular questions

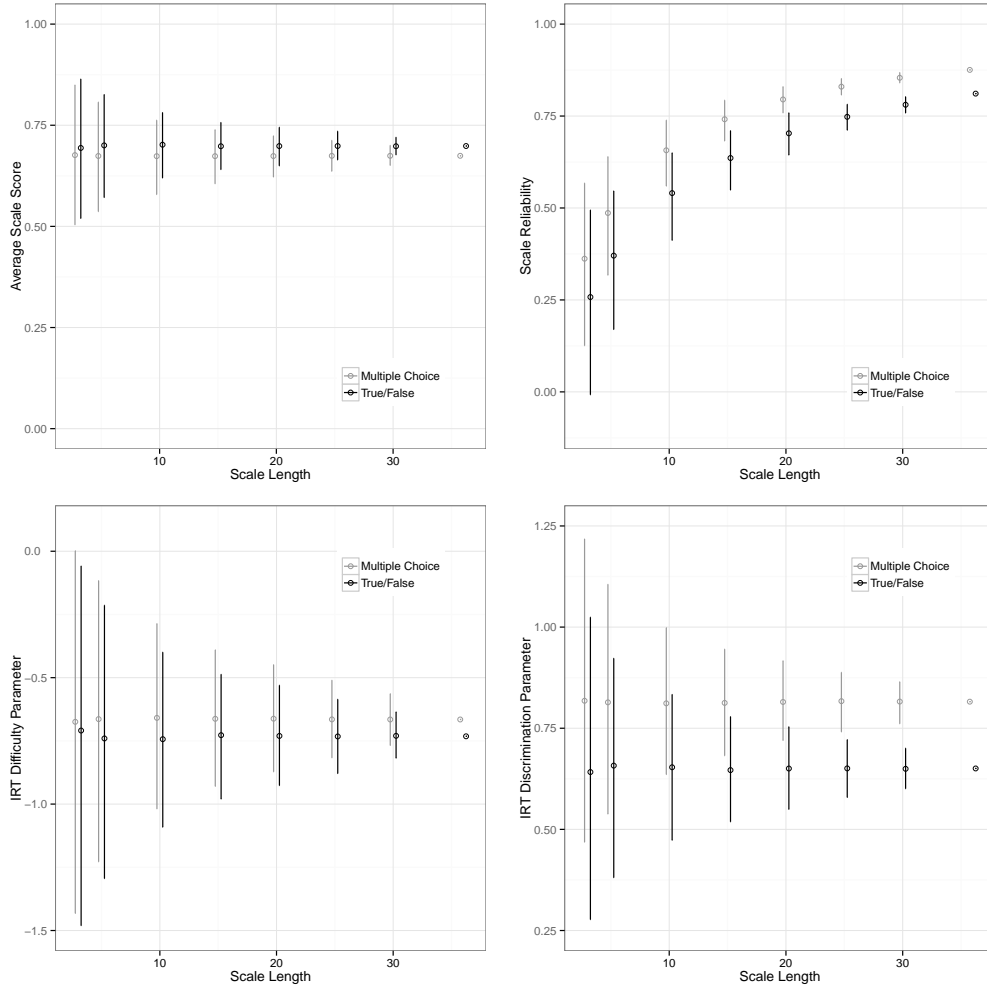


Figure OA-4: Scale Length, Question Type, and Scaling Measures [Study 1]
 NOTE: 95% Confidence Intervals are bootstrapped from a sample of 1000 draws of that scale length from our original 36 political items.

OA-2 Verbatim Question Wordings – Amazon Mechanical Turk Study

OA-2.1 True/False Questions

- civ.tf.1 For a bill to become law, the House of Representatives, but not the Senate, must approve its final wording. (FALSE - 72.4% Correct)
- civ.tf.2 The First Amendment to the Constitution protects the freedom of the press. (TRUE - 68.4% Correct)
- civ.tf.3 The Secretary of the Army is a position in the President's cabinet. (FALSE - 70% Correct)
- civ.tf.4 The judicial branch of the federal government determines whether or not laws passed by Congress are constitutional. (TRUE - 89.7% Correct)
- civ.tf.5 Fiscal policy is a country's central bank increasing or decreasing the money supply to control inflation. (FALSE - 36.3% Correct)
- civ.tf.6 A U.S. Senator is elected for a term of six years. (TRUE - 79.7% Correct)
- peo.tf.1 Samuel Alito is the most recent Supreme Court justice to leave the court due to his death in February 2016. (FALSE - 76.9% Correct)
- peo.tf.2 John Kerry is the current U.S. Secretary of State (TRUE - 82.1% Correct)
- peo.tf.3 Ted Cruz is a U.S. Senator from Florida. (FALSE - 59.3% Correct)
- peo.tf.4 Loretta Lynch is the current U.S. Attorney General. (TRUE - 59.2% Correct)
- peo.tf.5 In recent years, Anthony Kennedy has been the Supreme Court justice who has most often been the swing vote in closely divided court cases. (TRUE - 55.2% Correct)
- peo.tf.6 Jack Lew is the current Chair of the Federal Reserve. (FALSE - 67.0% Correct)
- dom.tf.1 The Democratic party currently controls the US Senate, while the Republican party controls the US House. (FALSE - 62.9% Correct)
- dom.tf.2 To comply with the Affordable Care Act (health care reform), most Americans need to indicate they have health insurance coverage when they file their taxes. (TRUE - 92.5% Correct)
- dom.tf.3 The Keystone XL is a controversial proposed pipeline that would carry oil in large quantities from Canada to Texas. (TRUE - 80.2% Correct)
- dom.tf.4 In 2015, some members of Congress threatened a government shutdown to prevent the funding of the Internal Revenue Service (IRS). (FALSE - 53.3% Correct)
- dom.tf.5 In 2015, Congress passed bipartisan legislation to ban soaps and cosmetics with microbeads. (TRUE - 55.2% Correct)
- dom.tf.6 The proposed bill that would provide a path to citizenship for undocumented immigrants brought to the U.S. as children is called the HOPE Act. (FALSE - 38.4% Correct)
- int.tf.1 In November 2015, terrorist attacks in London killed 130 people. (FALSE - 62.4% Correct)
- int.tf.2 Russia is not a member of the European Union (EU). (TRUE - 75.5% Correct)
- int.tf.3 Bashar al-Assad is the current President of Egypt. (FALSE - 51.6% Correct)
- int.tf.4 An outbreak of zika virus in Latin America is responsible for a rise in birth defects. (TRUE - 80.7% Correct)
- int.tf.5 In mid-2015 and early 2016, stock market crashes in Brazil were responsible for global economic slumps. (FALSE - 79.3% Correct)
- int.tf.6 Iran is a predominately Muslim country with more Shia Muslims than Sunni Muslims. (TRUE - 82.6% Correct)

- mas.tf.1 Protests in New York first brought widespread attention to the Black Lives Matter movement. (FALSE - 56.8% Correct)
- mas.tf.2 It is legal in all 50 U.S. states for gays and lesbians to marry. (TRUE - 64.3% Correct)
- mas.tf.3 People can legally buy marijuana (and other cannabis products) for recreational use in Colorado. (TRUE - 96.2% Correct)
- mas.tf.4 An anti-government militia recently seized and occupied federal land in Idaho for over a month. (FALSE - 55.9% Correct)
- mas.tf.5 In 2015, police arrested over 170 people in Fresno, CA after a motorcycle gang shootout killed nine. (FALSE - 53.1% Correct)
- mas.tf.6 The U.S. Department of Justice is investigating Chicago's police department following the public release of a video of an officer shooting a 17-year-old. (TRUE - 81.0% Correct)
- cel.tf.1 The Emmy Awards are for Broadway performances. (FALSE - 82.7% Correct)
- cel.tf.2 American Idol features contestants in a singing competition. (TRUE - 97.7% Correct)
- cel.tf.3 Brad Pitt is currently married to Jennifer Aniston. (FALSE - 86.9% Correct)
- cel.tf.4 Jon Hamm is the award-winning star of Mad Men. (TRUE - 67.9% Correct)
- cel.tf.5 The movie soundtrack to The Bodyguard is the best-selling soundtrack of all time. (TRUE - 43.2% Correct)
- cel.tf.6 Drake's album Purpose was the best-selling album of 2015. (FALSE - 76.1% Correct)

OA-2.2 Multiple Choice Questions

NOTE: Correct answers are preceded with a *. Response option order was randomized for respondents, except on questions where there is a natural numerical order (e.g., civ.m.6). The raw percentage of responses is displayed next to each response option.

- civ.m.1 For a bill to become a law, who must approve its final wording?
 - The Senate (17.1%)
 - The House of Representatives (3.6%)
 - The Senate and the House of Representatives (11.9%)
 - * The Senate, the House of Representatives, and the President (67.4%)
- civ.m.2 The First Amendment to the Constitution protects which of these rights?
 - * Freedom of the press (71.5%)
 - Right to vote (10.9%)
 - Right to a speedy trial (3.6%)
 - Right to bear arms (14.0%)
- civ.m.3 Which of the following is NOT a position in the president's cabinet?
 - * Secretary of the Army (71.4%)
 - Secretary of State (6.8%)
 - Secretary of the Treasury (2.6%)
 - Secretary of Agriculture (19.3%)
- civ.m.4 Which branch of the federal government determines whether or not laws passed by Congress are constitutional?
 - * Judicial branch (70.5%)
 - Legislative branch (18.1%)
 - Executive branch (7.3%)

- None of the above (4.1%)
- civ.m.5 What is fiscal policy?
 - * The government using taxes and spending to influence the economy (66.3%)
 - A country’s central bank increasing or decreasing the money supply to control inflation (25.9%)
 - The government using embargos and sanctions as part of foreign policy (3.1%)
 - The laws governing courts and trials (4.7%)
- civ.m.6 What is the term, in years, of a U.S. senator?
 - 2 years (24.6%)
 - 4 years (17.8%)
 - * 6 years (52.4%)
 - 8 years (5.2%)
- peo.m.1 Which Supreme Court justice died in February 2016?
 - Anthony Kennedy (4.7%)
 - * Antonin Scalia (83.3%)
 - Samuel Alito (6.8%)
 - Ruth Bader Ginsburg (5.2%)
- peo.m.2 Who is currently the U.S. Secretary of State?
 - * John Kerry (61.5%)
 - Hillary Clinton (23.4%)
 - Nancy Pelosi (7.8%)
 - Joe Lieberman (7.3%)
- peo.m.3 Which of these is a U.S. Senator from Florida?
 - * Marco Rubio (51.3%)
 - Ted Cruz (14.5%)
 - Jeb Bush (28.5%)
 - Rand Paul (5.7%)
- peo.m.4 Who is currently the U.S. Attorney General?
 - * Loretta Lynch (40.1%)
 - Eric Holder (42.2%)
 - Steny Hoyer (8.3%)
 - Elena Kagan (9.4%)
- peo.m.5 In recent years, which of the following Supreme Court justices has most often been the swingvote in closely divided court cases?
 - * Anthony Kennedy (38.9%)
 - Ruth Bader Ginsburg (19.7%)
 - Clarence Thomas (16.6%)
 - Antonin Scalia (24.9%)
- peo.m.6 Who is the Chair of the Federal Reserve?
 - * Janet Yellen (48.2%)
 - Paul Ryan (15.0%)

- Jack Lew (20.2%)
- Susan Collins (16.6%)
- dom.m.1 What party currently controls the House and Senate in the US Congress?
 - * The Republicans control both (59.8%)
 - The Democrats control both (3.1%)
 - The Democrats control the Senate, the Republicans control the House (22.7%)
 - The Republicans control the Senate, the Democrats control the House (14.4%)
- dom.m.2 To comply with the Affordable Care Act (health care reform), most Americans need to indicate they have health insurance coverage when they
 - * File their taxes (87.1%)
 - Change their address (3.1%)
 - Receive a driver’s license (4.1%)
 - Vote in an election (5.7%)
- dom.m.3 What is the name of the controversial proposed pipeline that would carry oil in large quantities from Canada to Texas?
 - * Keystone XL (80.9%)
 - Exxon XL (8.8%)
 - Continental XL (6.2%)
 - Alberta XL (4.1%)
- dom.m.4 In 2015, some members of Congress threatened a government shutdown to prevent the funding of which organization?
 - * Planned Parenthood (82.0%)
 - The Internal Revenue Service (IRS) (4.6%)
 - The Environmental Protection Agency (EPA) (9.8%)
 - The U.S. Department of Health (3.6%)
- dom.m.5 What household item did Congress ban in 2015 with bipartisan support?
 - * Soaps and cosmetics with microbeads (46.9%)
 - Plastic bags (6.7%)
 - E-cigarettes (5.7%)
 - Cold medicines with pseudoephedrine (40.7%)
- dom.m.6 What is the name of the proposed bill that would provide a path to citizenship for undocumented immigrants brought to the U.S. as children?
 - * DREAM Act (62.7%)
 - HOPE Act (19.7%)
 - FAIR Act (16.1%)
 - LEADER Act (1.6%)
- int.m.1 In which European city did terrorist attacks kill 130 people in November 2015?
 - * Paris (92.2%)
 - Munich (2.1%)
 - London (4.7%)
 - Rome (1.0%)
- int.m.2 Which of these countries is not a member of the European Union (EU)?

- * Russia (63.2%)
- Greece (10.4%)
- Germany (3.6%)
- Bulgaria (22.8%)
- int.m.3 Who is Bashar al-Assad?
 - * President of Syria (66.5%)
 - Leader of ISIS (19.1%)
 - Leader of al Qaeda (5.7%)
 - President of Egypt (8.8%)
- int.m.4 An outbreak of which disease in Latin America is responsible for a rise in birth defects?
 - * Zika virus (89.0%)
 - Hanta virus (2.6%)
 - Ebola (4.7%)
 - Smallpox (3.7%)
- int.m.5 Which country's stock market crashed in mid-2015 and early 2016, causing global economic slumps?
 - * China (60.3%)
 - Japan (11.3%)
 - Germany (15.5%)
 - Brazil (12.9%)
- int.m.6 In which of the following predominately Muslim countries are there more Shia Muslims than Sunni Muslims?
 - * Iran (43.5%)
 - Saudi Arabia (30.6%)
 - Afghanistan (19.7%)
 - Indonesia (6.2%)
- mas.m.1 In which U.S. state did protests first bring widespread attention to the Black Lives Matter movement?
 - * Missouri (77.7%)
 - California (3.6%)
 - New York (13.0%)
 - Indiana (5.7%)
- mas.m.2 In how many U.S. states are gay and lesbian couples legally allowed to marry?
 - 0 (0.5%)
 - 22 (17.6%)
 - 36 (22.8%)
 - * 50 (59.1%)
- mas.m.3 People can legally buy marijuana (and other cannabis products) for recreational use in which of these states?
 - * Colorado (89.1%)
 - California (6.7%)
 - Wisconsin (2.1%)
 - Vermont (2.1%)
- mas.m.4 In which U.S. state did members of an anti-government militia recently seize and occupy federal land for over a month?

- * Oregon (71.9%)
 - Montana (8.3%)
 - Idaho (9.4%)
 - Texas (10.4%)
- mas.m.5 In which U.S. city in 2015 did police arrest over 170 people after a motorcycle gang shootout that killed nine?
 - * Waco, TX (45.6%)
 - San Bernardino, CA (15.0%)
 - Fresno, CA (9.3%)
 - Galveston, TX (30.1%)
- mas.m.6 The U.S. Department of Justice is investigating which city's police department following the public release of a video of an officer shooting a 17-year-old?
 - * Chicago (51.3%)
 - Los Angeles (6.8%)
 - St. Louis (31.9%)
 - Atlanta (9.9%)
- cel.m.1 Which of these awards is for Broadway performances?
 - * Tony Awards (83.9%)
 - Emmy Awards (4.7%)
 - Academy Awards (8.8%)
 - Grammy Awards (2.6%)
- cel.m.2 Which of these talents are featured on the show American Idol?
 - * Singing (96.9%)
 - Cooking (1.5%)
 - Dancing (1.5%)
 - Fashion design (0.0%)
- cel.m.3 Brad Pitt is currently married to which person?
 - * Angelina Jolie (92.3%)
 - Jennifer Aniston (5.7%)
 - Gwyneth Paltrow (1.5%)
 - Juliette Lewis (0.5%)
- cel.m.4 Jon Hamm is the award-winning star of which television series?
 - * Mad Men (58.2%)
 - Game of Thrones (17.5%)
 - Breaking Bad (13.9%)
 - Homeland (10.3%)
- cel.m.5 Which movie soundtrack is the best-selling of all time?
 - * The Bodyguard (17.1%)
 - Saturday Night Fever (25.9%)
 - Titanic (43.5%)
 - Footloose (13.5%)
- cel.m.6 Which musical artist's album Purpose was the third best-selling album of 2015?
 - * Justin Bieber (33.2%)
 - One Direction (20.7%)
 - Taylor Swift (29.5%)
 - Drake (16.6%)

OA-2.3 Party Placement Question

Which party is generally more supportive of...?
[Democratic Party, Republican Party]

- Raising taxes on high incomes (Democratic Party - 90.3% Correct)
- Passing laws to improve the social and economic position of blacks (Democratic Party - 93.7% Correct)
- Restricting abortion (Republican Party - 92.0% Correct)
- Deporting undocumented immigrants (Republican Party - 91.0% Correct)
- Increasing the use of wind and solar power (Democratic Party - 87.8% Correct)
- Reducing defense spending (Democratic Party - 87.0% Correct)

OA-3 Verbatim Question Wordings – IGS Poll Study

OA-3.1 Delli-Carpini & Keeter Battery

- (Question 1 - Easy) Which of the following statements accurately describes the U.S. House of Representatives?
 - The Democratic Party controls 90% of seats and the Republican Party controls 10% (5.1%)
 - The Democratic Party controls 57% of seats and the Republican Party controls 43% (25.6%)
 - * The Democratic Party controls 43% of seats and the Republican Party controls 57% (63.9%)
 - The Democratic Party controls 10% of seats and the Republican Party controls 90% (5.4%)
- (Question 1 - Hard) Which of the following statements accurately describes the U.S. House of Representatives?
 - The Democratic Party controls 62% of seats and the Republican Party controls 38% (10.4%)
 - The Democratic Party controls 55% of seats and the Republican Party controls 45% (21.7%)
 - The Democratic Party controls 48% of seats and the Republican Party controls 52% (28.7%)
 - * The Democratic Party controls 43% of seats and the Republican Party controls 57% (39.2%)
- (Question 2 - Easy) Who is the current Vice President of the United States?
 - * Joe Biden (90.9%)
 - Bill Clinton (0.4%)
 - Barack Obama (6.6%)
 - Dick Cheney (2.1%)
- (Question 2 - Hard) Who is the current Vice President of the United States?
 - * Joe Biden (94.1%)
 - Harry Reid (1.0%)
 - John Kerry (3.7%)
 - John Roberts (1.3%)
- (Question 3 - Easy) Which of these statements best describes the American political parties?
 - * The Democratic Party is liberal and the Republican Party is conservative (84.0%)
 - The Democratic Party is conservative and the Republican Party is liberal (8.6%)
 - Both parties are liberal (3.4%)
 - Both parties are conservative (4.0%)
- (Question 3 - Hard) Which of these statements best describes the American political parties?

- * The Democratic Party is liberal on social and economic issues, while the Republican party is conservative on social and economic issues (75.6%)
- The Democratic Party is conservative on social and economic issues, while the Republican Party is liberal on social and economic issues (9.2%)
- The Democratic Party is liberal on social issues and conservative on economic issues, while the Republican Party is conservative on social issues and liberal on economic issues (10.9%)
- The Democratic Party is conservative on social issues and liberal on economic issues, while the Republican Party is liberal on social issues and conservative on economic issues (4.4%)
- (Question 4 - Easy) Who in government is responsible for judicial review?
 - * The Supreme Court (81.5%)
 - Congress (14.2%)
 - The President (3.2%)
 - The Department of Agriculture (1.0%)
- (Question 4 - Hard) Who in government is responsible for judicial review?
 - * The Supreme Court (74.1%)
 - The House of Representatives (6.3%)
 - The Senate (8.3%)
 - The Attorney General (11.3%)
- (Question 5 - Easy) How large of a congressional majority is necessary to override a presidential veto?
 - 50% (8.2%)
 - * 67% (75.9%)
 - 90% (11.1%)
 - 100% (4.9%)
- (Question 5 - Hard) How large of a congressional majority is necessary to override a presidential veto?
 - 50% (6.1%)
 - 60% (25.5%)
 - * 67% (44.1%)
 - 75% (24.3%)

OA-3.2 Our Battery

- (Question 1 - Easy) Which of these countries is not a member of the European Union (EU)?
 - * Russia (84.7%)
 - France (2.4%)
 - Germany (3.5%)
 - Belgium (9.4%)
- (Question 1 - Hard) Which of these countries is not a member of the European Union (EU)?
 - * Russia (72.3%)
 - Greece (12.7%)
 - Germany (1.9%)
 - Bulgaria (13.2%)
- (Question 2 - Easy) Who is currently the U.S. Attorney General?
 - * Loretta Lynch (64.7%)

- Barack Obama (3.3%)
- John Ashcroft (20.9%)
- Madeline Albright (11.0%)
- (Question 2 - Hard) Who is currently the U.S. Attorney General?
 - * Loretta Lynch (65.4%)
 - Eric Holder (22.4%)
 - Steney Hoyer (5.1%)
 - Elena Kagan (7.1%)
- (Question 3 - Easy) In which U.S. state did protests first bring widespread attention to the Black Lives Matter movement?
 - * Missouri (56.8%)
 - California (13.7%)
 - New York (19.4%)
 - Indiana (10.1%)
- (Question 3 - Hard) In which U.S. state did protests first bring widespread attention to the Black Lives Matter movement?
 - * Missouri (46.2%)
 - Illinois (14.5%)
 - Maryland (15.5%)
 - Mississippi (23.9%)
- (Question 4 - Easy) In how many U.S. states are gay and lesbian couples legally allowed to marry?
 - 0 (2.2%)
 - 22 (45.4%)
 - 36 (18.9%)
 - * 50 (33.6%)
- (Question 4 - Hard) In how many U.S. states are gay and lesbian couples legally allowed to marry?
 - 35 (31.5%)
 - 36 (29.8%)
 - 48 (4.3%)
 - * 50 (34.4%)
- (Question 5 - Easy) The First Amendment to the Constitution protects which of these rights?
 - * Freedom of the press (67.7%)
 - Right to vote (16.2%)
 - Right to a speedy trial (2.5%)
 - Right to bear arms (13.6%)
- (Question 5 - Hard) The First Amendment to the Constitution protects which of these rights?
 - * Freedom of the press (59.9%)
 - Right to vote (11.2%)
 - Privacy (6.4%)
 - Equal protection under the law (22.6%)

OA-4 Verbatim Question Wordings—CA Field Poll Study

- (Q1 - version 1) How large of a congressional majority is necessary to override a presidential veto?
 - 1/2 (Half) (7.3%)
 - 3/5 (Three-fifths) (11.2%)
 - * 2/3 (Two-thirds) (67.3%)
 - 3/4 (Three-quarters) (13.7%)
- (Q1 - version 2) A two-thirds majority is necessary in Congress for which of the following?
 - Passing a bill in conference committee (14.7%)
 - Ending a filibuster (6.2%)
 - * Overriding a presidential veto (66.4%)
 - Passing budget legislation (12.8%)
- (Q1 - version 3) How large of a congressional majority is necessary to override a presidential veto?
 - 50% (5.4%)
 - 60% (17.4%)
 - * 67% (54.4%)
 - 75% (22.8%)
- (Q2 - version 1) The First Amendment to the U.S. Constitution protects which of these rights?
 - * Freedom of the press (73.8%)
 - The right to vote (12.8%)
 - The right to bear arms (9.3%)
 - The right to a speedy trial (3.1%)
- (Q2 - version 2) Which of these amendments to the U.S. Constitution protects freedom of the press?
 - * First (70.0%)
 - Second (11.3%)
 - Sixth (9.7%)
 - Fifteenth (6.5%)
- (Q3 - version 1) Who in the U.S. government is responsible for judicial review?
 - The House of Representatives (6.1%)
 - The Senate (11.8%)
 - * The Supreme Court (69.3%)
 - The Attorney General (10.1%)
- (Q3 - version 2) Which of these jobs in government is the U.S. Supreme Court responsible for?
 - Writing legislation (5.6%)
 - Negotiating treaties (2.0%)
 - * Judicial review (76.3%)
 - Enforcing laws (15.8%)
- (Q4 - version 1) Who is currently the U.S. Attorney General?
 - John Ashcroft (11.5%)
 - Eric Holder (14.2%)
 - * Loretta Lynch (68.9%)

- Elena Kagan (3.4%)
- (Q4 - version 2) What is Loretta Lynch’s current job in the U.S. government?
 - Supreme Court Justice (14.8%)
 - * Attorney General (73.0%)
 - FBI Director (2.3%)
 - Ambassador to the U.N. (9.2%)
- (Q5 - version 1) Which of the following are Americans legally allowed to do in all 50 U.S. states? (NOTE: The administration of this item mistakenly asked respondents yes/no for each response option, so the percentages do not sum to 100%)
 - Use marijuana for medical purposes (12.5%)
 - Openly carry a firearm in public (12.1%)
 - Commit adultery (38.0%)
 - * Marry someone of the same sex (47.9%)
- (Q5 - version 2) In how many U.S. states are Americans legally allowed to marry someone of the same sex?
 - 0 (3.1%)
 - 22 (31.5%)
 - 36 (22.7%)
 - * 50 (41.4%)
- (Q6 - version 1) The two candidates running for California’s open U.S. Senate seat this year are Kamala Harris and Loretta Sanchez. What are these candidates’ party affiliations?
 - Harris is a Democrat and Sanchez is a Republican (6.3%)
 - Harris is a Republican and Sanchez is a Democrat (13.6%)
 - * Harris and Sanchez are both Democrats (77.7%)
 - Harris and Sanchez are both Republicans (1.5%)
- (Q6 - version 2) The two candidates running for California’s open U.S. Senate seat this year are Kamala Harris and Loretta Sanchez. Kamala Harris is a Democrat. What party does Loretta Sanchez belong to?
 - * Democratic (83.7%)
 - Republican (12.8%)
 - Libertarian (0.5%)
 - Green (1.5%)
- (Q6 - version 3) The two candidates running for California’s open U.S. Senate seat this year are Kamala Harris and Loretta Sanchez. Loretta Sanchez is a Democrat. What party does Kamala Harris belong to?
 - * Democratic (86.4%)
 - Republican (8.6%)
 - Libertarian (1.0%)
 - Green (2.0%)

OA-5 Construction of Batteries in Figure 2

OA-5.1 Optimal Battery, Multiple Choice

- Party placement: “Increasing the use of wind and solar power” (Democratic)
- Civics: “What is the term, in years, of a U.S. senator?” (6 years)
- Party placement: “Deporting undocumented immigrants” (Republican)
- Mass politics/social movements: “In how many U.S. states are gay and lesbian couples allowed to marry?” (50)
- Mass politics/social movements: “In which U.S. city in 2015 did police arrest over 170 people after a motorcycle gang shootout that killed nine?” (Waco, TX)

OA-5.2 Optimal Battery, True/False

- Party placement: “Increasing the use of wind and solar power” (Democratic)
- Political figures: “Loretta Lynch is currently the U.S. Attorney General” (T)
- People: “Jack Lew is the current Chair of the Federal Reserve” (F)
- Civics: “Fiscal policy is a country’s central bank increasing or decreasing the money supply to control inflation” (F)
- Mass politics/social movements: “People can legally buy marijuana (and other cannabis products) for recreational use in Colorado” (T)²²

OA-5.3 Too-Easy Battery, Multiple Choice

- International politics/events: “In which European city did terrorist attacks kill 130 people in November 2015?” (Paris)
- Mass politics/social movements: “People can legally buy marijuana (and other cannabis products) for recreational use in which of these states?” (Colorado)
- International politics/events: “An outbreak of which disease in Latin America is responsible for a rise in birth defects?” (Zika virus)
- Domestic policy: “What is the name of the controversial proposed pipeline that would carry oil in large quantities from Canada to Texas?” (Keystone XL)
- Political figures: “Which Supreme Court justice died in February 2016?” (Antonin Scalia)

OA-5.4 Too-Easy Battery, True/False

- Mass politics/social movements: “People can legally buy marijuana (and other cannabis products) for recreational use in Colorado” (T)
- Domestic policy: “To comply with the Affordable Care Act (health care reform), most Americans need to indicate they have health insurance coverage when they file their taxes” (T)
- Civics: “The judicial branch of the federal government determines whether or not laws passed by Congress are constitutional” (T)
- Domestic policy: “The Keystone XL is a controversial proposed pipeline that would carry oil in large quantities from Canada to Texas” (T)
- Political figures: “John Kerry is the current U.S. Secretary of State” (T)

²²Those who would like to know more about the characteristics of Mechanical Turk respondents may be interested to know that this was by far the least difficult political knowledge item on our survey. 96% of respondents answered this question correctly. This is comparable to the 97% of correct responses to the true statement that *American Idol* is a singing competition, the least difficult question overall.

OA-5.5 Too-Hard Battery, Multiple Choice

- Political figures: “In recent years, which of the following Supreme Court justices has most often been the swing vote in closely divided court cases?” (Anthony Kennedy)
- Political figures: “Who is currently the U.S. Attorney General?” (Loretta Lynch)
- International politics/events: “In which of the following predominately Muslim countries are there more Shia Muslims than Sunni Muslims?” (Iran)
- Mass politics/social movements: “In which U.S. city in 2015 did police arrest over 170 people after a motorcycle gang shootout that killed nine?” (Waco, TX)
- Domestic policy: “What household item did Congress ban in 2015 with bipartisan support?” (Soaps and cosmetics with microbeads)

OA-5.6 Too-Hard Battery, True/False

- Civics: “Fiscal policy is a country’s central bank increasing or decreasing the money supply to control inflation” (F)
- Domestic policy: “The HOPE Act is a proposed bill that would provide a path to citizenship for undocumented immigrants brought to the U.S. as children” (F)
- International politics/events: “Bashar al-Assad is the current President of Egypt” (F)
- Domestic policy: “In 2015, some members of Congress threatened a government shutdown to prevent the funding of the Internal Revenue Service (IRS)” (F)
- Mass politics/social movements: “In 2015, police arrested over 170 people in Fresno, CA after a motorcycle gang shootout killed nine” (F)