Advance Access publication July 17, 2013

Political Analysis (2013) 21:393–406
doi:10.1093/pan/mpt010

Harder Than It Looks: Coding Political Knowledge on the ANES

Matthew DeBell

Institute for Research in the Social Sciences, Stanford University,
30 Alta Road, Stanford, CA 94305
e-mail: debell@stanford.edu

Edited by R. Michael Alvarez

Political knowledge research faces a problem, perhaps even a crisis. For two decades, the American National Election Studies asked open-ended questions about political knowledge and coded answers using procedures that are neither reliable nor replicable and that were never shown to be optimally valid. Consequently, conclusions based on these widely used measures of the public's competence are in doubt. This article presents several new and overdue methodological improvements: coding knowledge data using formal and specific coding rules based on a substantive rationale for the validity of the codes, recognizing partially correct answers, using multiple coders working independently, using machine coding, and testing reliability and validity. The new methods are an improvement because they are transparent and replicable and they produce valid and extremely reliable knowledge data. Further, machine coding produces codes nearly identical to those from a team of human coders, at much lower cost.

1 Introduction

A new set of open-ended political knowledge questions debuted on the American National Election Studies (ANES) Time Series survey in 1986. These questions asked "what job or political office" is now held by various prominent officials. The questions were repeated on subsequent ANES studies and in recent years have regularly named the Vice President, the Chief Justice of the United States, the Speaker of the US House of Representatives, and the Prime Minister of the United Kingdom. At times, other offices were used, including Secretary of State and Senate Majority Leader. Answers were coded as correct or incorrect by ANES staff or University of Michigan Survey Research Center staff, and the codes were released on the public-use ANES data files. Quite a bit of literature has used these variables in the quarter century since their debut (e.g., Karp 1995; Delli Carpini and Keeter 1996; Mondak 2000; Barabas 2002; Brewer 2003; Ladd 2007; Luskin and Bullock 2011). The questions have been a model for similar questions on other surveys, but the ANES remains a leading provider of such data.

Optimal practices for coding open-ended responses to questions such as these include the following requirements intended to promote the transparency and validity of research methods (cf. Krosnick et al. 2008; Lupia 2008):

- Coding rules are substantively correct.
- Codes are tested for criterion validity.
- A substantive rationale is articulated for the construct validity of the coding rules.

Author's note: I thank Catherine Wilson, Lucila Figueroa, and Mackenzie Israel-Trummel for accurately coding thousands of data points, Skip Lupia and Jon Krosnick for leading by example and for their helpful comments on an earlier draft of this paper, and Catherine Warrick and anonymous reviewers for more recent feedback. For materials to replicate the statistical analysis in this article, see DeBell (2013). Supplementary materials for this article are available on the *Political Analysis* Web site.

© The Author 2013. Published by Oxford University Press on behalf of the Society for Political Methodology.
All rights reserved. For Permissions, please email: journals.permissions@oup.com

https://doi.org/10.1093/pan/mpt010 Published online by Cambridge University Press

- Coding rules allow nuanced codes to reflect partially correct answers.
- Coding rules are formal, specific, and comprehensive.
- Coding rules are documented and publicly disclosed.
- Coding rules and methods are consistent over time so that statistics are comparable.
- Source data (open-ended responses) are publicly disclosed.
- Coding is performed by coders working with records of the open-ended responses, rather than on the fly by interviewers during the interview.
- Coding is performed by multiple coders, with inter-coder reliability reported.

Prior to the ANES 2008 Time Series study, *none* of these practices were consistently followed. Problems in previous coding practices came to light because of Gibson and Caldiera's (2009) investigation of public knowledge of the Supreme Court and were documented by Krosnick et al. (2008). For example, in 2004, instructions said that "Prime Minister of the United Kingdom" is not the correct answer for the office held by Tony Blair, though that answer was (obviously!) correct. No peer-reviewed criterion validity tests have been published. Codes have always used dichotomous values for correct or incorrect answers, making it impossible to discriminate between completely incorrect answers and answers that are correct but not as specific as an ideal answer (e.g., saying that William Rehnquist was on the Supreme Court, but not that he was Chief Justice). Coding instructions have variously been vague or undocumented or both. No written records of instructions exist for coding done between 1986 and 1998. Between 2000 and 2004, some coding criteria changed. Coding has relied on open-ended data that were never released to the public before 2008. In 2004, coding was performed by interviewers during the interview, and no systematic record of the actual answers was made. Typically, the coding was performed by just one person, and reliability has never been reported for any ANES knowledge codes.

These weaknesses mean we have no way of knowing that the knowledge codes produced from 1986 through 2004 are accurate or consistent. The codes have been described as "so problematic that we conclude that they support only the simplest and grossest inferences" (Gibson and Caldiera 2009, 432). We also have no blueprint for optimally coding these answers in the future. This has created a data quality crisis, with analysts unable to proceed with desired research until the coding methods are fixed (Luskin and Bullock 2011). ANES needs a new and better coding system for open-ended political knowledge questions—a transparent and replicable system (Lupia 2008)—so that analysts can believe the results and benefit from more reliable and valid data.

The preceding observations point to an opportunity to substantially improve ANES practices, of course, but the opportunity is much broader. Survey research practitioners more generally have not placed a high priority on rigorously handling open-ended data for simple questions, even on high-visibility, big-budget studies. Attempting to find publicly released open-ended responses, inter-coder reliabilities, criterion validity analysis, and comparisons of alternate coding schemes for data from other major infrastructure surveys is largely an exercise in frustration. This sort of work has not been under the scrutiny of peer review. Clearly it needs to be, for it took more than twenty years for these problems with political knowledge to come to light.

This article fixes these problems with the political knowledge data. In the initial sections, I describe a new coding process using multiple coders working independently as well as a computer algorithm to automatically implement practically identical codes. In the later sections, I evaluate the codes produced by this new process. I argue that the codes should be evaluated in terms of inter-coder reliability, scale reliability, criterion validity as expected main-effect associations with other variables, and criterion validity as expected moderating effects in relationships between other variables. By all of these tests, the new method gives reliable and valid data. The article concludes with recommendations for further improvement of political knowledge measures and of open-ended coding of such questions in political science.

2 Coding Schemes

A "coding scheme" is a set of rules for how text should be described with numbers that represent the text's content. This section discusses such rules.

https://doi.org/10.1093/pan/mpt010 Published online by Cambridge University Press

2.1 *Lessons from Prior Work*

One of the weaknesses of prior coding practices has been the use of only one set of dichotomous coding criteria in which answers were coded as correct or incorrect based on fairly demanding standards. For example, the Chief Justice of the United States, more popularly known as the Chief Justice of the Supreme Court, is one of nine judges on the highest federal court, yet answers identifying this person as "a judge" or as "on the supreme court" were marked wrong (Mondak 2001). An important disadvantage of a single dichotomous coding scheme such as this one is that it cannot discriminate between answers that are entirely incorrect and answers that are correct but not thorough. Moreover, depending on the research questions with which particular investigators approach the data, one coding scheme may be better than another. Therefore, the research community should have more than one set of codes that describe the responses so that data users can choose the codes best suited to their particular research questions.

Delli Carpini and Keeter (1996) and subsequent researchers (e.g., Mondak 2000) distinguish respondents who are misinformed, uninformed, partially informed, and fully informed. An optimal coding scheme would be flexible enough to allow analysts to distinguish all these types of responses. To that end, Martinez and Craig (2010) developed a coding scheme that ranked answers to all four office identification questions as clearly correct, leadership, accurate, party only, ballpark, or incorrect, with separate codes for do not know and refused. This represents the only completed systematic attempt of which I am aware to develop a comprehensive coding scheme for the ANES office recall questions.[1] The key instructions were as follows (and the entire coding descriptions are in supplementary appendix 1):

- *Clearly correct*: responses identify the exact leadership position held by the person, though they may also include some additional extraneous correct or incorrect information. . . .
- *Leadership*: responses do not cite the specific office held by the political figure, but identify her or him as a political leader in the appropriate branch of government or as a party leader.
- *Accurate*: responses identify the branch or locale of the officeholder, but fail to specifically note his or her leadership role.
- *Party only*: responses only name the party the figure belongs to or, in a few cases, mention a political ally of the figure without any reference to an office that the person holds.
- *Ballpark*: answers here indicate some leadership role or policy with which the figure is associated, but the specific office is lacking or misplaced and the response is technically inaccurate.

Martinez and Craig should be commended for transparency in their coding methodology, because they have published both their code descriptions and the codes themselves. As they note, classification judgments can be difficult. For example, consider how a few answers, shown in Table 1, were scored for the Pelosi and Cheney questions.

Reasonable researchers can disagree about these codes. Applying Martinez and Craig's rules, "ranking senator" identifies Pelosi as a leader in the legislative branch and so might be coded as "leader" instead of "incorrect." One might prefer to say that "senate majority leader" qualifies for the "leader" code just as well as "democrat majority leader of the senate" or "she's head of something in Washington." The latter response identifies neither Pelosi's branch of government nor a party leadership position, so the rules do not seem to warrant the "leader" code. For Cheney, one answer mentioned "secretary of state" but was not coded like the other answers naming the same office, and one said only "secretary, not sure," which might be too vague to qualify for the ballpark code. Moreover, the "secretary of state" responses (coded as "ballpark") might be interpreted as meeting the "leadership" coding criteria because they identify him "as a political leader in the appropriate branch of government"—assuming the Vice President is in the executive branch, which has been a point of contention (Subhawong 2008).

[1] Jon Krosnick and Arthur Lupia have announced at ANES public meetings that they intend to produce comprehensive new coding schemes for ANES open-ended questions, but results of this effort were not reported before this article was completed.

https://doi.org/10.1093/pan/mpt010 Published online by Cambridge University Press

Table 1 Examples of judgment challenges in coding

	Martinez and Craig code
Pelosi response	
Ranking senator	Incorrect
Congresswoman in Ca. 8th district	Accurate
In charge of the senate floor	Ballpark
Senate majority leader	Ballpark
Democratic head of the senate	Ballpark
Chairman of congress she presides over it	Ballpark
Democrat majority leader of the senate	Leader
Big wig in congress	Leader
She's head of something in Washington	Leader
Head of house	Leader
Cheney response	
Senate	Incorrect
Chairholder to the senate	Incorrect
I don't know. Republican senator	Incorrect
I know the name but not office//can't think.secretary of state	Incorrect
Secretary of state	Ballpark
Secretary, not sure	Ballpark
Chancellor of foreign affairs	Ballpark

Martinez and Craig acknowledge the difficulty but do not quantify how often such difficulties might lead different researchers to different decisions. These difficulties are inherent in most coding schemes, which require coders to exercise a moderate degree of judgment.

Apart from the judgment calls about application of the rules, one may question the rules themselves. If the rules give "ballpark" to a claim that Dick Cheney was "chancellor of foreign affairs," one might find this too generous. And if one does not find this too generous, one might want to give some ballpark credit to "chairholder of the senate," since the Vice President is the President of the Senate and Cheney did preside over the Senate and cast tiebreaking votes. One might say that calling Pelosi "senate majority leader" is just plain wrong. One might say "head of house" is a substantially better answer than "big wig in congress" and "democrat majority leader of the senate," because the "big wig" answer is vague and does not specify the chamber, and the "senate" answer specifies the wrong chamber and the wrong office. These difficulties are a persistent conundrum of coding partial credit for the office recall questions.

2.2 *New Coding Rules*

These observations reinforce two conclusions. First, it is important to calculate and report inter-rater reliability so that we are transparent about the extent to which subjective judgments by particular coders affect the data. When reliability is extremely high, a reader can be more confident that he or she would reach the same conclusions as the coders. When reliability is not high or is not reported, we have no such confidence, and the review of Table 1 shows why.

Second, it is important to develop and disclose coding rules that are as precise and comprehensive as possible. This does not eliminate subjectivity from the codes. Instead, it *moves* subjectivity from behind the curtain of the coder's mind into the sunshine of peer review, in publicly disclosed precise rules for the coding. Once transparency is achieved through disclosure of highly specific rules, the subjective judgments that were the basis of those rules can be scrutinized and improved upon. This is impossible when subjectivity is bound up in the undocumented process by which coders make difficult judgment calls on their own. With transparent and comprehensive rules, the judgments that drive the coding are in the rules themselves and can be readily evaluated and, hopefully, revised for the better by other scholars.

https://doi.org/10.1093/pan/mpt010 Published online by Cambridge University Press

One of the primary goals for the present project was to write rules that are clear enough to achieve extremely high inter-rater reliability. Skeptics' attention can then be fruitfully aimed at the rules they see instead of the coding process they cannot see.

To these ends, code development proceeded in two stages. The first product was a conceptual ranking of response quality for the knowledge items, and the second was a set of codes that would allow researchers to map responses to these response quality ranks but be flexible enough to be mapped to other indices of quality based on different judgments about what it is meaningful for the public to know.

The conceptual ranking of answer quality places the office title first, a description of the titular office second, and a looser but accurate description of the position third. The rationale for this approach is that it favors more specific and demanding but meaningful reports of knowledge. For the Speaker, the ideal answer would name the office, a lesser answer would indicate a leadership position in the House, and a minimally useful answer would identify her as a member of Congress. For the Prime Minister of the United Kingdom, the ideal answer would name the office and the country, a lesser answer would identify him as a national leader of the United Kingdom, and a minimally useful answer would identify his nationality or identify his office without specifying the country. For the Vice President, knowledge of the office is so discrete and the actual answers have so little variance that the only answer it seems necessary to recognize is "Vice President" (or an abbreviation or nickname such as "VP"). For the Chief Justice, the ideal answer would name the office, a lesser answer would identify him as the head of the Supreme Court, and a minimally useful answer would identify him as a judge or a member of the Supreme Court.

The codes and associated coding rules were developed to fit three criteria: (1) to differentiate the types of responses just described, (2) to allow analysts to test alternative methods of converting codes to scores, and (3) to minimize the amount of coder judgment required and therefore maximize inter-coder reliability. The coding instructions are in supplementary appendix 2. The instructions are designed to promote inter-rater reliability by being as brief, simple, and specific as possible. For most codes, the tenor of the instructions is that of an algorithm rather than a guideline. For instance, Pelosi scheme 1 calls for identifying the simultaneous presence of the words "house" and "speaker," and nothing else matters. This sort of task can be completed with greater reliability than coding rules that call for nuanced judgment.

Coding decisions are generally dichotomous, but by assigning more than one code per answer, codes allow researchers to account for partially correct answers. There are twelve substantive coding variables for the four questions: three codes for the Speaker, three for the Prime Minister, five for the Chief Justice,[2] and one for the Vice President.

3 Data

Data were from the ANES 2008 Time Series study, including redacted open-ended responses. The survey was representative of adult US citizens. It was conducted using face-to-face interviews in two waves, immediately before and after the November 2008 election. Knowledge questions were asked on the postelection wave, which had 2102 respondents and an estimated overall response rate of 58%. Lupia et al. (2009) gives methodological details. The data are at www.electionstudies.org. In this article, parameter estimates are weighted and use design-consistent calculations of sampling errors (DeBell 2010).

4 Coding Procedure

Coding was performed independently by four coders: myself, a second social scientist with a doctorate, and two doctoral students. I developed the coding rules. I had prior experience coding open-ended survey data, but the other coders had none.

[2]John Roberts became Chief Justice in 2005. Another John Roberts was the chief White House correspondent for CBS from 1999 to 2006 and was with CNN at the time of the 2008 survey. The dozen respondents in 2008 who identified Roberts as a journalist were, therefore, correct. I created a separate variable to identify answers that identified Roberts this way, but the answer is so rare that doing so makes little difference and those results are not discussed here.

https://doi.org/10.1093/pan/mpt010 Published online by Cambridge University Press

Each coder individually made 25,224 coding decisions for the twelve substantive coding variables, a process that averaged approximately 18 hours per coder. This is several codes per minute, or approximately 2.5 s per code, which is fast but practical because many responses are short, the coding rules are simple, and the coding decisions under most of the coding rules can be made almost instantly. An additional 8408 decisions were made by each coder to identify the "don't know and nonresponse" items, for a total of 31,530 decisions per coder.[3] Collectively, we assigned 126,120 codes.

After independently completing the coding, we assembled a list of every code assignment for which the four coders were not unanimous in their independent decisions and we met to resolve these differences. There were 182 cases of disagreement (where a case is one respondent's answer to one knowledge question, and may include one or more disagreements among the coders) out of the 8408 answers that were coded; coders were unanimous in their coding decisions for 98% of the respondents' answers. After meeting, of the 25,224 substantive codes we assigned, we reached unanimity for 25,221 answers and did not reach unanimity for the remaining three. For the 8408 answer/nonresponse codes, we reached unanimity for all but five. The nonunanimous cases are shown in supplementary appendix 3.

Review and discussion of the initially nonunanimous decisions revealed three sources of differences: careless mistakes, differing interpretations of the coding rules, and differing interpretations of interviewer misspelling. Careless mistakes were coding differences where the correct code was obvious to all coders, yet a decision was initially incorrect. This occurred, for example, six times in 8408 coding decisions for the Vice President item. Interpretation of the coding rules was usually not difficult, but some clarifications to the rules were made (shown in italics in supplementary appendix 2) to cure ambiguities. For example, in Pelosi scheme 2, it was unclear how much latitude to give to "the response means Speaker," so this was corrected to "if the response means Speaker or head or leader." Interviewer spelling caused problems by making it unclear what the respondent actually said. For example, coders were divided on the interpretation of "hold a conseaa seat," with some preferring to credit "conseaa seat" as meaning "Congress seat" and others disagreeing.[4]

Final codes were assigned in three ways. The *consensus* code was the code that all four coders independently assigned when they did so unanimously, or the code the four coders agreed upon in conference when they reviewed all cases where their independent decisions were not initially unanimous. For the few cases of nonunanimity, a plurality vote was used. I treat the consensus codes as the "correct" or at least the best available decisions and use them as the benchmark for evaluating individuals' decisions and other codes. The *modal* or plurality code was the decision made most often by the independent coders. By comparing the results from the consensus and modal codes, we can assess the value of the conference. Finally, as described in the next section, a *script* code was produced by a computer algorithm.

5 Computer Coding Algorithms

Machine coding can reduce labor costs and make code assignments perfectly reliable in the sense that any researcher can apply the algorithm to the data and produce identical results. It can also be completed in moments instead of days. Moreover, because computers implement algorithms in a rote fashion, developing coding rules for computers to follow has the virtue of demanding that the researcher formalize all coding rules.

[3]Respondents may be resistant to guessing because they prefer appearing uninformed to appearing misinformed or because guessing requires mental effort and saying "don't know" is easier. I created a variable to distinguish answers from nonresponse to allow researchers to attend to these issues. In practice, this primarily means distinguishing "don't know" answers from answers that include a guess. Interviewers were trained to "probe" "don't know" answers to encourage a guess, so every "don't know" answer constitutes a refusal. Therefore, all refusals were collapsed into the no answer category.

[4]ANES has audio recordings of most of these interviews, so it would be possible for ANES staff to review and correct at least some of these errors.

https://doi.org/10.1093/pan/mpt010 Published online by Cambridge University Press

Machine coding of ANES open-ended data is experimental. There have been forays (Fan 2008; Skalski 2008; Esuli and Sebastiani 2010), but without validity tests or comparison to other methods. To compare results to human coding, I developed coding algorithms for each coding scheme for each knowledge question. The syntax is in supplementary appendix 4. It aims to make the same code assignments as humans who follow the instructions in supplementary appendix 2.

6 Analytic Approach

The rest of this article evaluates the quality of the new knowledge codes and considers whether it is likely to matter if researchers choose one or another. Before presenting the results I outline the evaluation methods and criteria.

Good data are reliable and valid. Reliability means consistency: repeating a procedure should produce the same result. Validity means correctness: the variable actually measures what it is supposed to measure. To evaluate the quality of the knowledge codes, I examine inter-coder reliability, scale reliability, and criterion validity. The aim is to discover whether the measures are reliable and valid, and if so, which coding approach maximizes these properties.

Inter-coder reliability, also known as inter-rater reliability, is a measure of the reproducibility of the coding judgments. High reliability indicates that codes are being applied consistently and that independent coders agree; imperfect reliability means errors cloud the data. Several measures of reliability are available. The rate (percent) of agreement between coders has the virtue of being easy to interpret, but it does not account for the rate of chance agreement and can therefore be misleading. Cohen's kappa (κ), a widely used conservative measure of inter-coder agreement that is not inflated by chance agreement (Cohen 1960), is therefore also helpful. Kappa gives the proportion of agreement between raters after the agreement attributable to chance has been removed from both the numerator and denominator. Some researchers have interpreted kappa values above 0.75 as excellent (Fleiss, Levin, and Palik 2003) and others have interpreted values above 0.80 as acceptable (Krippendorff 2004). The goal is reliability approaching 100% and $\kappa = 1.0$.

Scale reliability is important when multiple items are interpreted as measuring the same thing and are combined into one scale variable. Analysts routinely interpret knowledge questions as measuring a single latent construct of general political knowledge (though this is often a problematic practice; Lupia 2006). The reliability of such scales is commonly measured with Cronbach's alpha, and I do so here for all possible combinations of the knowledge codes for a four-item knowledge scale. There are no firmly established standards for interpreting Cronbach's alpha, but values below about 0.6 are usually judged to imply the items do not consistently measure a single latent construct.

I examine the criterion validity of the recall scores by testing their association with four other measures of political knowledge: correctly identifying the party with a majority of seats in the US Senate, identifying the majority party in the US House of Representatives, correctly placing the Democratic Party to the left of the Republican Party on a left-right ideological placement scale, and correctly placing Barack Obama to the left of his Republican opponent in the 2008 campaign for president, John McCain. The rationale for these validity tests is the assumption that, for example, the ability to recall Nancy Pelosi's office should be related to knowing the majority party in the House. This does not reflect an assumption that the office recall questions are good measures of general political knowledge, which they probably are not (Lupia 2006). Nor does it reflect an assumption that both sets of knowledge items measure the same latent trait of general political knowledge. Rather, it reflects the tame assumption that knowledge of these particular facts is correlated.

The reliability and validity analyses presented below use the codes that the four coders reached in consensus.[5] To reflect the potential for the coding schemes to be combined into a scale, I also

[5]In the interest of space, only results from the consensus codes will be shown. Results from the mode and script codes are essentially the same, which is to be expected because both alternate methods of producing final codes yielded nearly identical results most of the time.

https://doi.org/10.1093/pan/mpt010 Published online by Cambridge University Press

computed four-point scales for Pelosi, Brown, and Roberts, respectively, reflecting the conceptual ranking of response quality discussed earlier. The Pelosi variable is coded as follows: Speaker of the House 1, head of Congress 0.67, member of Congress 0.33, not correct 0. Brown: Prime Minister of the United Kingdom 1, leader of the United Kingdom 0.67, foreign leader or Briton 0.33, not correct 0. Roberts: Chief Justice of the Supreme Court or Chief Justice of the United States 1, head of the high court 0.67, judge 0.33, not correct 0.[6]

7 Descriptive Results

The different coding schemes produce somewhat different results, but all versions indicate relatively small proportions of the public were able to identify the offices held by these people. Table 2 shows the proportions.

Cheney's office was identifiable by 73% of Americans.

Brown's office was identifiable by 5% when held to the standard that the respondent must name the office as "Prime Minister" and name the country as the United Kingdom, Britain, or England. This number rose to 6% (not a statistically significant difference from 5) when the standard was relaxed to naming Brown as a national leader of the same country, and to 7%

Table 2 Percentage correctly identifying office

Person	*Standard for correctness*	*Percent*	*SE*
Pelosi			
Scheme 1	House Speaker	37	2.3
Scheme 3	Leader of House or Congress	38	2.4
Scheme 2	House, Congress, or Rep.	48	2.4
Cheney			
Scheme 1	Vice President	73	1.8
Brown			
Scheme 1	PM of UK/GB/England	5	1.0
Scheme 2	Leader of UK/GB/England	6	1.0
Scheme 3	Nationality or foreign leader	7	1.0
Roberts			
Scheme 1	Chief Justice of US or SC	4	0.6
Scheme 2	Head of high court	5	0.7
Scheme 3	"chief" + "justice"	5	0.8
Scheme 4	Judge	16	1.4
Scheme 5	Judge	16	1.4
$N = 2102$			

Note. For comparison, the 2004 ANES found 9% correctly identified Dennis Hastert's office as Speaker of the House, 84% identified Dick Cheney's as Vice President, 62% identified Tony Blair's office as Prime Minister of the United Kingdom (or England), and 28% identified William Rehnquist's office as Chief Justice. Given the different times and people, it is impossible to know how much of the differences are attributable to coding methods or to actual differences in knowledge.

[6]To estimate how much credit the partially correct answers deserve, I used a regression method similar to that of Mondak and Davis (2001). I developed a knowledge criterion score based on the sum of correct answers on the other available knowledge variables: knowing left-right placement of Obama and McCain, left-right placement of the parties, the Senate majority party, the House majority party, and Cheney as Vice President. Then I used dummy variables for full credit and partial credit answers to predict this score. The resulting estimates could be used to determine the value of partially correct answers relative to fully credited answers. However, due to small sample sizes, the confidence intervals for the estimates of partial credit were too large to reliably differentiate the partial credit estimates from one another or from default (equidistant interval) values in all cases. Lacking an empirical basis to specify intervals for these codes, I followed a default practice of assigning equal intervals. In the future, more data should enable accurate estimation of these intervals.

https://doi.org/10.1093/pan/mpt010 Published online by Cambridge University Press

(again, nonsignificant) when the standard was relaxed to naming Brown as a foreign leader or as from the United Kingdom, Britain, or England.

Pelosi's office was identifiable by 37% of Americans when held to the standard that the respondent must name the office as "Speaker of the House." This was 38% (not statistically significant) when the standard was relaxed to "leader" of the House or Congress, and 48% when the standard was relaxed to say that she was a Representative or a member of the House or of Congress. (It would be higher if "Senator" were accepted as correct.)

Roberts' office was identifiable by just 4% of Americans when held to the standard that the respondent must name the office as "Chief Justice" of the United States or of the Supreme Court. This was 5% when the standard allowed the respondent to name the position as "head" or "leader" or similar synonym for Chief and also allowed alternative wording for "Supreme Court." It was also 5% when the coding criteria were simply that the words "chief" and "justice" both be present, which was the de facto coding rule used by ANES in 2000 and was very nearly the operative rule in 1998 and 1996 (Krosnick et al. 2008). When the standard was relaxed to saying Roberts was a judge, the proportion correct rose to 16%.

8 Inter-coder Reliability

Inter-coder reliabilities were extremely high for most codes and were perfect in many cases. Table 3 shows these reliabilities. In each case, the reliability is calculated with reference to the consensus codes. Reliability for the Roberts coding schemes ranged from 100% to 98.6% ($\kappa = 1.00$–0.939) for the human coders and 99.7–99.0 for the computer script. Reliability for Pelosi ranged from 99.9% to 98.4% for humans and from 99.0% to 93.4% for the computer. Reliability for Brown ranged from 100% to 99.0% for humans and was 99.8% for the computer. Reliability for Cheney ranged

Table 3 Inter-rater reliabilities, compared to consensus codes

	Roberts scheme 1		*Roberts scheme 2*		*Roberts scheme 3*		*Roberts scheme 4*	
Code	*Percent*	*Kappa*	*Percent*	*Kappa*	*Percent*	*Kappa*	*Percent*	*Kappa*
Coder 1	100.0	1.000	100.0	1.000	100.0	1.000	99.8	0.992
Coder 2	99.9	0.985	100.0	0.994	99.9	0.982	99.8	0.990
Coder 3	99.9	0.985	99.8	0.967	100.0	1.000	99.6	0.984
Coder 4	100.0	1.000	100.0	0.994	100.0	1.000	99.0	0.960
Script	99.7	0.954	99.6	0.940	99.7	0.961	99.1	0.963
Mode	100.0	1.000	100.0	1.000	100.0	1.000	100.0	1.000
	Roberts scheme 5		*Pelosi scheme 1*		*Pelosi scheme 2*		*Pelosi scheme 3*	
Coder 1	99.8	0.990	99.9	0.997	99.3	0.988	98.5	0.965
Coder 2	99.6	0.981	99.9	0.998	99.6	0.992	98.4	0.964
Coder 3	99.7	0.988	99.8	0.996	99.4	0.990	98.4	0.963
Coder 4	98.6	0.939	99.4	0.987	98.6	0.975	99.0	0.976
Script	99.0	0.959	99.0	0.975	93.4	0.876	98.6	0.967
Mode	99.9	0.994	100.0	0.999	99.5	0.998	98.6	0.968
	Brown scheme 1		*Brown scheme 2*		*Brown scheme 3*		*Cheney scheme 1*	
Coder 1	100.0	1.000	100.0	1.000	99.9	0.991	99.8	0.996
Coder 2	100.0	1.000	99.9	0.990	99.9	0.987	100.0	1.000
Coder 3	100.0	1.000	100.0	1.000	99.8	0.979	99.9	0.997
Coder 4	99.9	0.984	99.0	0.898	99.6	0.966	99.9	0.997
Script	99.8	0.972	99.8	0.974	99.8	0.983	99.3	0.984
Mode	100.0	1.000	100.0	1.000	100.0	1.000	100.0	1.000

Note. Percents shown as 100.0 are rounded in some cases.

https://doi.org/10.1093/pan/mpt010 Published online by Cambridge University Press

from 100% to 99.8% for humans and was 99.3% for the computer. Overall, reliability was very high, demonstrating that independent coders consistently applied the coding rules—which has never before been shown for ANES knowledge items.

The lowest computer reliability was for Pelosi scheme 2, where spelling errors by the interviewers hobbled the algorithm's performance. In all other cases, the algorithm's reliability rounded to 99% or 100%.

The reliability of the modal (plurality) answer was 100% in most cases, meaning that the consensus reached by the four coders for each code was usually the same code that was assigned by the most coders. This is a good result because it shows that discussion among the coders rarely led in a different direction from coders' independent judgments.

9 Scale Reliability

Table 4 displays reliabilities (Cronbach's alpha) of scales composed of the four knowledge items coded using alternate combinations of the codes. There are forty-five possible combinations of these codes that can build a four-item knowledge scale, and all are shown in Table 4. The scale reliabilities range from about 0.50 to 0.65. The better reliabilities are obtained with Roberts scheme 5 and Pelosi scheme 2 or 3, all of which reflect partial credit. Even the better reliabilities are at best moderate, which suggests that combining the knowledge items into a single knowledge variable may not be warranted.

Table 4 Scale reliabilities for knowledge scales

Scheme set	*Alpha*	*Scheme set*	*Alpha*
111	0.56	123	0.52
211	0.56	223	0.53
311	0.57	323	0.53
411	0.56	423	0.61
511	0.63	523	0.62
112	0.55	131	0.55
212	0.56	231	0.56
312	0.56	331	0.56
412	0.56	431	0.56
512	0.63	531	0.63
113	0.57	132	0.55
213	0.58	232	0.56
313	0.58	332	0.56
413	0.57	432	0.56
513	0.65	532	0.63
121	0.50	133	0.57
221	0.50	233	0.58
321	0.50	333	0.58
421	0.60	433	0.57
521	0.60	533	0.65
122	0.50		
222	0.50		
322	0.51		
422	0.60		
522	0.60		

Note. Scheme set digits are Roberts (hundreds digit), Pelosi (tens), Brown (ones). For example, 523 indicates Roberts scheme 5, Pelosi scheme 2, Brown scheme 3. Cheney scheme 1 is included for all.

https://doi.org/10.1093/pan/mpt010 Published online by Cambridge University Press

Table 5 Criterion validity of office recall using other knowledge measures

	Knowledge criterion variables, logistic regression coefficients							
	House majority		*Senate majority*		*Party placement*		*Candidate placement*	
Office recall	*Coeff.*	*SE*	*Coeff.*	*SE*	*Coeff.*	*SE*	*Coeff.*	*SE*
Pelosi								
Scheme 1	1.79	0.14	1.42	0.12	1.40	0.14	1.62	0.13
Scheme 2	1.90	0.14	1.51	0.13	1.47	0.14	1.67	0.14
Scheme 3	1.82	0.13	1.46	0.12	1.41	0.14	1.63	0.13
Scale	1.92	0.14	1.53	0.13	1.49	0.14	1.71	0.14
Cheney								
Scheme 1	1.26	0.14	1.20	0.13	1.10	0.14	1.41	0.14
Brown								
Scheme 1	1.85	0.27	1.33	0.27	2.26	0.51	2.22	0.58
Scheme 2	1.84	0.26	1.28	0.27	1.98	0.52	1.98	0.55
Scheme 3	1.89	0.25	1.29	0.25	1.66	0.38	1.69	0.40
Scale	1.99	0.27	1.37	0.27	2.08	0.47	2.08	0.52
Roberts								
Scheme 1	1.99	0.27	1.37	0.33	1.86	0.50	1.97	0.48
Scheme 2	1.89	0.34	1.40	0.30	2.02	0.49	2.13	0.47
Scheme 3	1.94	0.27	1.48	0.30	1.88	0.45	1.98	0.43
Scheme 4	2.41	0.25	1.96	0.24	2.80	0.37	2.77	0.32
Scheme 5	2.46	0.29	2.00	0.26	2.79	0.40	2.77	0.34
Scale	2.89	0.40	2.17	0.34	3.45	0.67	3.50	0.59

Note. All coefficients are statistically significant at $p < 0.05$. All variables are coded to range from 0 to 1. $n = 2102$ for each regression.

10 Validity

I used logistic regression to predict correct answers to four criterion variables (candidate left-right placement, party placement, House majority party, and Senate majority party) using the various office recall codes one at a time. Table 5 presents the logistic regression coefficients from sixty logistic regressions in which office recall is the independent variable and the knowledge criterion variables are dependent. All results are statistically significant ($p < 0.05$).

For Pelosi, there is little variation in the coefficients among the coding schemes. The four-point scale is more strongly associated with the criterion knowledge variables than the individual coding schemes are, but none of the coefficients are statistically significantly different from one another. All, however, are greater than zero, which is an indication of validity.

For Cheney, all criterion variables are significantly predicted.

For Brown, as with Pelosi, all predictors are significant, but none differ from one another.

For Roberts, there is a consistent pattern that the more relaxed coding schemes (4 and 5) are more strongly associated with the criterion variables than the stricter coding schemes (1 and 3). The strongest association is with the four-item scale.

These results show that the office recall codes pass a test of criterion validity, because all of the office recall coding schemes are associated with other knowledge items. These tests also show that these coding schemes produce similar results for Pelosi and Brown. For Roberts, however, there is variation in the quality of the data from the different coding schemes, with the more relaxed schemes producing greater validity. Better validity from the more relaxed and nuanced coding schemes shows that the old practice of scoring the Chief Justice item strictly was needlessly throwing away useful information, as observed by Gibson and Caldeira (2009).

I also tested validity by examining the association between the office recall items and other correlates of political sophistication, and with voter turnout, and as a moderator of policy

https://doi.org/10.1093/pan/mpt010 Published online by Cambridge University Press

preferences' effect on vote choice. In the interest of space, these results are shown in supplementary appendix 5. They retell the story that knowledge coded with these methods passes the validity test of producing expected associations, and that particularly for the Chief Justice item, giving partial credit improves the measurement of knowledge.

11 Discussion and Recommendations

Scholars have awaited improved ANES knowledge codes (e.g., Luskin and Bullock 2011). This article has reported one of the first attempts to develop rigorous coding methods for political knowledge items that have been widely used and are now twenty-five years old. The present findings support recommendations about how to code political knowledge and about how analysts should use and interpret the codes.

By using simple and carefully specified coding instructions, we achieved very high reliability for all coding schemes tested. Results with the current group of coders seem likely to generalize to other trained and supervised coders who are verbally fluent and attentive to detail, because the procedure is algorithmic and depends on care and verbal fluency rather than expert judgment. The experimental computer script demonstrated excellent correspondence with the codes developed by four coders, matching those consensus codes for about 99%–100% of cases in eleven of the twelve coding schemes. Machine coding shows considerable promise for these data, but so does the work of humans, provided we have clear and simple written instructions.

The main impediment to accurate machine coding for the scheme where machine coding's reliability was a relatively low 93% ($\kappa = 0.88$) was the interviewers' spelling. It is simply awful. The reliability and perhaps the validity of the data could be improved if interviewers transcribed respondents' answers with fewer spelling errors. Integrating spell-check into the CAPI software, and having interviewers take a few moments immediately after each interview to correct their typing, probably would help improve accuracy. Otherwise, to handle these problems, it is appropriate to have a human coder check the work of a machine script. The particularly good news is that machine coding of these open-ended data is fast, inexpensive, quite reliable, and fairly easy.

As noted above, Gibson and Caldeira (2009) concluded that the ANES open-ended knowledge data were "so problematic that . . . they support only the simplest and grossest inferences." For the old codes of the Chief Justice item, this seems true. But if the 2008 data are an indication, it is possible that prior years' codes of other office recall questions were not so bad. For each question, coding scheme 1 approximates the rule that ANES appears to have used in the past, and scheme 1's codes proved reliable and valid by the multiple measures applied here. We will know how the reliability and validity of the old items compares to codes produced under a new system such as the one presented here if and when the new system is applied to old ANES data. This cannot be done with publicly available data now. To make it possible, ANES should redact any sensitive material from the open-ended answers in its restricted-access archive and make those redacted files public, as it has done with the 2008 data.

Validity tests against other measures of political knowledge (and correlates of political sophistication, turnout prediction, and the moderating effect of policy preferences on vote choice, shown in supplementary appendix 5) suggest the validity of knowledge measures tends to be better (and is never worse) when nondichotomous measures of knowledge are used for these items. The assignment of partial credit for Roberts, in particular, is important for optimizing the predictive and criterion validity of this variable. For the other variables, assigning partial credit is less important for modeling public opinion but perhaps no less important for accurately gauging the public's overall familiarity with these officials. Future codes, therefore, should be assigned on a continuum of quality. In the present system, analysts relying on these measures will have reliable and valid data if they use any of these coding schemes, and will probably optimize their choices for most purposes if they use any of the Pelosi schemes, use the Cheney scheme, use the Brown scheme 3 or the scale built from the Brown codes, and use the Roberts scheme 4, 5 or the scale built from the Roberts codes.

Beyond these details, this research emphasizes the importance of transparency as a guiding principle for research. The lack of transparency in past practices regarding political knowledge

https://doi.org/10.1093/pan/mpt010 Published online by Cambridge University Press

led to problems going unrecognized for many years. The transparent and replicable procedures outlined here demonstrate that these problems are solvable and that knowledge codes are reliable and valid when we follow best practices: formal written rules with a substantive rationale, multiple coders working independently (or a computer algorithm given clean data), and partial credit where partial credit is due.

Funding

This project was supported by grant SES-0937715 from the National Science Foundation to Stanford University (Gary Segura and Simon Jackman, Principal Investigators); conclusions are the author's alone.

References

Abramowitz, Alan. 1995. It's abortion, stupid: Policy voting in the 1992 presidential election. *Journal of Politics* 57:176–86.

Barabas, Jason. 2002. Another look at the measurement of political knowledge. *Political Analysis* 10:209–22.

Brewer, Paul R. 2003. Values, political knowledge, and public opinion about gay rights. *Public Opinion Quarterly* 63:173–201.

Carmines, Edward G., and James A. Stimson. 1980. The two faces of issue voting. *American Political Science Review* 74:78–91.

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20:37–46.

DeBell, Matthew. 2010. *How to analyze ANES survey data*. ANES Technical Report Series, no. nes012492. http://electionstudies.org/resources/papers/nes012492.pdf (accessed September 28, 2010).

———. 2013. Replication data for validation: Harder than it looks: Coding political knowledge on the ANES [dataset], http://hdl.handle.net/1901.1/21558. IQSS Dataverse Network [Distributor] V1 [Version].

Delli Carpini, Michael X., and Scott Keeter. 1996. *What Americans know about politics and why it matters*. New Haven, CT: Yale University Press.

Downs, Anthony. 1957. *An economic theory of democracy*. New York: Harper.

Esuli, Andrea, and Fabrizio Sebastiani. 2010. Machines that learn how to code open-ended survey data. *International Journal of Market Research* 52:775–800.

Fan, David P. 2008. Computer coding of 1992 ANES like/dislike and MIP responses. Presented at conference on Optimal Coding of Open-Ended Survey Data, Ann Arbor, MI.

Fleiss, Joseph L., Bruce Levin, and Myunghee Cho Palik. 2003. *Statistical methods for rates and proportions*. 3rd ed. New York: Wiley.

Gibson, James L., and Gregory A. Caldeira. 2009. Knowing the Supreme Court? A re-consideration of public ignorance of the high court. *Journal of Politics* 71:429–41.

Goren, Paul. 1997. Political expertise and issue voting in presidential elections. *Political Research Quarterly* 50:387–412.

Karp, Jeffrey A. 1995. Explaining support for legislative term limits. *Public Opinion Quarterly* 59:373–91.

Krippendorff, Klaus. 2004. *Content analysis: An introduction to its methodology*. Thousand Oaks, CA: Sage.

Krosnick, Jon A., Arthur Lupia, Matthew DeBell, and Darrell Donakowski. 2008. Problems with ANES questions measuring political knowledge. ANES memorandum. http://www.electionstudies.org/announce/newsltr/20080324PoliticalKnowledgeMemo.pdf (accessed March 24, 2008).

Ladd, Jonathan McDonald. 2007. Predispositions and public support for the president during the war on terrorism. *Public Opinion Quarterly* 71:511–38.

Lupia, Arthur. 2006. How elitism undermines the study of voter competence. *Critical Review* 18:217–32.

———. 2008. Procedural transparency and the credibility of election surveys. *Electoral Studies* 27:732–39.

Lupia, Arthur, Jon A. Krosnick, Pat Luevano, Matthew DeBell, and Darrell Donakowski. 2009. *User's guide to the ANES 2008 time series study*. http://electionstudies.org/studypages/2008prepost/2008prepost_UsersGuide.pdf (accessed May 11, 2009).

Luskin, Robert C., and John G. Bullock. 2011. "Don't know" means "don't know": DK responses and the public's level of political knowledge. *Journal of Politics* 73:547–57.

Martinez, Michael D., and Stephen C. Craig. 2010. Dick Cheney? Didn't he shoot somebody? Measuring political knowledge in the 2008 American National Election Study. Presented at the annual meeting of the American Political Science Association, Washington, DC.

Marquis, Lionel. 2010. Understanding political knowledge and its influence on voting preferences in the 2007 federal election. *Swiss Political Science Review* 16:425–56.

Mondak, Jeffery J. 2000. Reconsidering the measurement of political knowledge. *Political Analysis* 8:57–82.

———. 2001. Developing valid knowledge scales. *American Journal of Political Science* 45:224–38.

Mondak, Jeffery J., and Belinda Creel Davis. 2001. Asked and answered: Knowledge levels when we will not take "don't know" for an answer. *Political Behavior* 23:199–224.

https://doi.org/10.1093/pan/mpt010 Published online by Cambridge University Press

Nicholson, Stephen P., Adrian Pantoja, and Gary M. Segura. 2006. Political knowledge and issue voting among the Latino electorate. *Political Research Quarterly* 59:259–71.

Skalski, Paul. 2008. CATA options for the coding of open-ended survey data. Presented at conference on Optimal Coding of Open-Ended Survey Data, Ann Arbor, MI.

Subhawong, Aryn. 2008. A realistic look at the vice presidency: Why Dick Cheney is an "entity within the executive branch." *Saint Louis University Law Journal* 53:281–308.

https://doi.org/10.1093/pan/mpt010 Published online by Cambridge University Press