Online Appendix Materials for
Winter, Nicholas J. G., Adam G. Hughes, and Lynn M. Sanders.

"Online Coders, Open Codebooks:
New Opportunities for Content Analysis of Political Communication."

Political Science Research and Methods

A1. Implementation of the Coding Portal

We programmed a customized coding interface on our webserver. The data entry form was implemented in Limesurvey, an open-source alternative to Qualtrics (see http://www.limesurvey.org). The interface itself was programmed in PHP and JavaScript; we also developed a set of utilities in Python for creating and managing the HITs and the mTurk qualifications we required workers to earn. We wrote custom template code for Limesurvey to optimize the coding form, as well as code that connects our platform with Amazon's API (for online workers) and with our local coding portal (for our research assistants) so that selecting our HIT, watching the ad, entering coding decisions, and submitting the results was as simple as possible. We will make all of our code available to researchers who wish to implement their own coding systems and/or modify them for their. Due to the modular design of the system, it would be relatively straightforward to adapt the system to use, e.g., Qualtrics, for the data collection component or to interface with other online labor forces or with local research assistants, as we did for our RAs. The portal also included a back-end interface for tracking progress, downloading data, and approving the mTurk work. We approved every submission except in two instances where workers consistently and repeatedly coded 30-second ads in less than 20 seconds each, suggesting that they were not actually watching them.

A2. Additional Details about Coding Process

The coding was completed in several waves, which we combine for the analyses we present here. Several items were added in the second wave, including the presence of economic appeals, appeals to anger and disgust, and whether a flag appears. The first wave also included several additional items we do not include in our analyses. These include specific reference to the physical appearance of the candidate, such as

hair, makeup, or general good (or bad) looks, which was also adapted from Hayes' coding scheme for newspaper coverage of campaigns. These sorts of references never appeared in the first-wave ads, so the item was dropped from subsequent coding. In addition, we explored various approaches to measuring references to gender roles, none of which ended up being coded often enough to support inclusion in the analyses presented here. Finally, the first wave of coding included items that attempted—and failed—to measure a distinction between sociotropic and pocketbook economic appeals (Kinder and Kiewiet 1979, 1981). We discuss the fate of these items in the paper's conclusion.

We recruited workers by posting the qualification HIT with the title, "University of Virginia Political Cognition Lab: Campaign Ad video coding" and with a description that read: "Watch 30-second political advertisements and code them for the presence of various elements and themes. The required qualification can be earned immediately by taking a brief training survey and confirming that you can view the videos." We compensated workers minimally for this HIT ($0.03) because we did not want to incentivize workers with no interest in actual content analysis to complete it.

A3. Mechanical Turk Requester Ratings and Reputation and Our Coders

Workers have incentives to do good work because requesters pay only for work they approve, and because Amazon makes workers' overall approval rates available to requesters. Requesters can limit their tasks to workers with high approval rates; they also can require that workers reside in the United States, or that workers pass a test or complete some assigned task to earn a "custom qualification." Because workers are paid by the (approved) task, they have a strong incentive to work quickly and effectively.

We require our workers to have approval rates of at least 95 percent because Peer and colleagues (2016) and Hauser and Schwarz (2015) show that workers with these rates are quite attentive to the tasks they complete as subjects in academic research studies and provide high-quality data. We set the minimum number of task at 100 because Amazon does not report the actual approval rate for a worker until they have

completed 100 tasks: "to ensure that a new Worker's approval rate is unaffected by these statistically meaningless changes, if a Worker has submitted less than 100 assignments, the Worker's approval rate in the system is 100%." See the mTurk API reference: https://docs.aws.amazon.com/AWSMechTurk/latest/AWSMturkAPI/ApiReference_QualificationRequirementDataStructureArticle.html (accessed October 7, 2018).

Overall, 1,235 mTurk workers took our qualification survey; all but eight were classified as eligible for coding because they completed the background survey and were able to view the ads. Of these, 526 went on to code at least one ad. Workers each coded an average of 53 ads, though the distribution is highly skewed: many dropped out after coding a few ads, moderate numbers coded dozens, and a few coded hundreds (median 7; range 1–1,159; standard deviation of 121). Seventy-five individuals completed 80 percent of the 27,335 ads coded by online workers. Our mTurk workforce was reasonably diverse with respect to age, education, and income, though not representative of the American public as a whole. Appendix table A14 provides demographic comparisons between the mTurk coders and the American population—the mTurk coders are somewhat younger, more likely to be white, and have lower income. Twenty percent identified as Republicans and 45 percent as Democrats, and their political knowledge was higher than the American average: based on a standard political knowledge battery modeled on the American National Election Study, our median coder scored 0.875 on a zero-to-one scale. About three quarters reported conducting at least some content coding in the past.

While Amazon tracks the worker approval rate to provide requesters with information on worker reputation, there is no official system to give workers parallel information on requesters' reputations. Most tasks are relatively short, so workers can protect themselves to some extent by completing one or two and waiting to see that they are paid before completing more. A set of informal online tools and communities have sprung up to allow workers to rate requesters (e.g., TurkOpticon, https://turkopticon.ucsd.edu/), to exchange

information about good and bad HITs (e.g., Reddit's HITs Worth Turking For, https://www.reddit.com/r/HITsWorthTurkingFor/), and to organize collective action to improve conditions for mTurk workers (e.g., We are Dynamo, http://www.wearedynamo.org/). The latter has developed a set of Guidelines for Academic Requesters; that document is aimed mostly at those using mTurk to recruit research subjects (http://wiki.wearedynamo.org/index.php/Guidelines_for_Academic_Requesters). Finally, on the ethics of mTurk as a source for research subjects, see Marinova (2016); on the ethics of online labor markets more generally and for jobs of the sort we describe here, see Fort (2011), Adda (2013), Busarovs (2013), and Williamson (2016).

A4. Detailed item-level reliability statistics and additional measures of reliability

There are many statistics and alternatives for weighting disagreements for non-binary coding (Gwet 2014). Each statistic makes somewhat different assumptions about the rating process and each weighting differs in the relative penalty for smaller vs. larger disagreements when coding more than two categories. We use ordinal weights (Gwet 2014, eq. 3.5.1) for the three-category emotions and quadratic weights (Gwet 2014, eq. 3.2.5) for our continuous ideology coding. Our results do not change with other weighting schemes or statistics; this appendix replicates the reliability analysis using Conger's (1980) kappa. We use the Stata package `kappaetc` to calculate these statistics (Klein 2017).

Appendix tables A1 and A2 present the item-level reliability statistics and gains from meta-coder aggregation; these correspond to the summary information presented in tables 2 and 3 of the paper.

To facilitate the analysis of gains from additional coders (below, appendix A6), we develop a measure based on the root-mean-squared coding error (RMSE). Because we have multiple coders for each ad, we can calculate the "error" for each coding decision as the difference between it and the average of all the other decisions for that item in that ad. These errors can be summarized in various ways. To generate an item-level measure of reliability that parallels Krippendorff's alpha or kappa (though with the opposite sign), we take the

square root of the average squared coding error, aggregated across ads and coders. This RMSE directly measures variability across coders. It extends naturally from binary items to our continuous ideology measure, and can be interpreted as the within-ad standard deviation of coding across the multiple coders, expressed in the units of the underlying coding scale. Tables A3 and A4 present item level reliability statistics using this measure; these tables correspond to the information presented in tables A1 and A2, respectively.

Finally, tables A5 and A6 present reliability measured a third way, using Conger's kappa (1980), which is a generalization of Cohen's kappa for 3 or more raters (see also Gwet 2014).

A5. Gains from Multiple Coders

We explored the relative reductions in RMSE as we increase the number of coders used to form a single meta-coder. For this analysis, we continue to focus on the roughly 200 ads for which we have at least 18 individual coders, and measure the RMSE, relative to a single coder, for meta-coders ranging two to eight coders. We randomly select a set of eight of the coders for possible inclusion in a meta-coder; the remaining coders (at least 10) are used to calculate the "truth" for each coding decision on that ad. Then, we calculate the coding by a single coder (using the value provided by the first of the eight randomly-selected coders), and for meta-coders of size two through eight (by averaging the values provided by the first two coders, then the first three, and so on). We calculate the error for each of these meta-coders as the difference between that coding and the "truth" based on the 10 or so coders who are not included in any of the meta-coders. To smooth out noise that depends on which coders are selected to serve as part of a meta-coder and which to calculate the "truth" to which the meta-coders are compared, we repeat this process ten times using different random selections, and average the results together. Finally, we summarize the squared errors for each type of coding decision across all the ads, and take the square root to generate the RMSE.

Table A7 shows the decrease in RMSE relative to a single coder for meta-coders made up of between two and eight individual coders. Figure A1 presents this information, disaggregated by coding decision; the

figure suggests that the story is essentially the same across all types of coding items, which is what we would expect given the mathematics of aggregation. We harvest the gains to aggregation quickly, with sharply declining marginal benefit from additional coders. Simply averaging two coders decreases RMSE by more than 20 percent, on average. Adding a third coder improves reliability by an additional 10 percent, and a fourth by 6.5 percent. Additional coders yield progressively smaller gains. Thus, we can improve reliability substantially by employing more than one, but still relatively few coders per ad—perhaps four or five.

A6. Analysis of time spent coding each ad

Factors affecting the amount of time spent coding an ad

Although a full analysis of the factors that affect coding time is beyond the scope of this paper, we ran two simple models that explored the impact on coding time of coder-level (appendix table A8) and ad-level (table A9) information that we have at hand. The first model indicates that workers who coded more ads worked faster: the model results imply that those who coded 10 or fewer averaged 143 seconds per ad; this drops to 65 seconds per ad for those who coded 100 or more. Two things are at play here: some of the slowest coders simply dropped out and stop working for us; in addition, among those who continued to code there appears to be a learning curve, with coding speeding up a bit as they get practice. For example, among workers who coded at least 100 ads, they averaged 98 seconds on the first ten, and 68 seconds on ads 50 through 99.

There were a few other aggregate differences among coders: older coders were notably slower (e.g., those above age 50 averaged 32 seconds slower than those 25 and under), and partisans were about 15 seconds faster per ad than independents (14 seconds for Republicans; 17 for Democrats). Interestingly, coders' level of political knowledge did not systematically affect their coding speed.

Turning to ad-level characteristics, we lack much contextual information about the ads. From what we have, House and Senate ads took very similar time to code, as did ads that (according to wmp) focused on

policy, personal characteristics, or both. Compared with candidate-sponsored ads, party- and interest-group ads were very slightly faster to code, by 5 and 4 seconds, respectively. Positive ads were about 7 seconds faster than comparative ads, and about 5 seconds faster than attack ads.

We did not require workers to view the ad before they began to fill in the coding form. Our intuitions on this are mixed: on the one hand, such a requirement might encourage them to watch the ad with fuller attention. On the other, it would require them to remember any coding decisions they can make early on—for example, if they see the favored candidate holding a flag while talking about unemployment in the first seconds of the ad, they would have to remember three coding decisions until the end of the ad before they could click the appropriate buttons. As we discuss in the conclusion, this sort of question about the best procedures for coding is amenable to systematic empirical exploration using our approach.

Impact of time spent coding on reliability and validity

Analyzing reliability at the level of the individual coding decision presents a challenge, as inter-coder reliability is generally calculated at the level of the coding item: it is an aggregate property of a coding decision among a group of coders. Therefore, we turn to root-mean-squared error (RMSE) reliability measure developed in Appendix A4.

Focusing just on 30-second ads (which make up the vast majority of the data), we group the time spent coding into five categories, corresponding to the first three quartiles, the 75th-90th percentiles, and those above the 90th percentile. We regress decision-level disagreement on indicator variables for the type of coding decision (economic appeal, flag appearance, traits, etc.), indicators for time spent coding the ad, and the interactions among them. The results are displayed in table A10 and figure A2. Simply, there is no evidence of systematic differences in reliability by time spent. We interpret this to mean that, although different coders were faster or slower and different ads required more or less time to code, on the whole coders spent the time necessary to code each ad reliably.

For validity, we focus on favored candidate ideology, as that is the coding decision for which we have the best validity measure. We regress the individual coding decision about favored-candidate ideology on the DW-NOMINATE score for that candidate, interacted with the time spent on the ad. This regression coefficient is functionally equivalent to the correlation between coding and ideology (at the ad level) that we report in the main text, but has the advantage of allowing for easy interaction with time. The results, in table A11 and figure A3, indicate that there is no statistically-significant impact of time spent on validity. However, as coders spend more time on an ad, there is a steady—though small and insignificant—increase in our estimated validity.

More broadly, this analysis gives a very quick view of the sorts of analyses of coding quality that are possible when multiple measures of each coding decision are available.

A7. Coder learning or fatigue

We explored whether coders got systematically better (or worse) as they worked.

We might expect coders to improve as they learn from experience; on the other hand, they might get worse if they become more careless over time. In the results below, we found no evidence of large, systematic changes, though there was a hint that the reliability of trait and economic coding declined slightly after coders had seen 100 ads. Focusing on our best validity test—favored candidate ideology—there was no statistically-significant change with coder experience, though again there was a hint of a small dip after 100 ads. We take these findings to indicate that this is not a major concern, though researchers should take care to ensure that their most active coders remain vigilant.

To assess reliability rely on the RMSE measure that we can calculate at the level of the coding decision, as we did in Appendix A6. Because we have information on the date and time each ad was coded, we can calculate the cumulative number of ads coded by a particular worker at the moment that they complete each ad. We group all the coding decisions into a five categories: the first 10 ads encountered by a coder, ads 11-50,

51-100, 100-200, and 200+. (The results presented here are unaffected by different grouping schemes, and also in models that rely on the natural logarithm of the sequence number.)

In table A12 we regress decision-level disagreement on dummy variables for the type of coding decision (economic appeal, flag appearance, traits, etc.), indicators for the sequence grouping, and the interactions among them. The results are displayed in figure A4. There are no major changes, although there are small increases in error rate (i.e., decreases in reliability) for the identification of economic appeals and for traits. These increases are statistically significant only after 100 ads coded, and never very large substantively speaking. There are very slight decreases over time for economic tone and ideology, though these are not statistically significant.

For validity, we focus on favored candidate ideology, as that is the coding decision for which we have the best validity measure. We regress the individual coding decision about favored-candidate ideology on the DW-NOMINATE score for that candidate, interacted with coding sequence. This regression coefficient is functionally equivalent to the correlation between coding and ideology (at the ad level) that we report in the main text, but has the advantage of allowing for easy interaction with coding sequence. The results, presented in table A13 and figure A5, give no indication that validity changes systematically as coders gain more experience.

A8. The ambiguity of flags in (some) ads

Online appendix figures A6 through A10 show screenshots of five ads, typical of those that generated disagreement on the presence of an American flag. In the first two, a small flag appears *very* briefly: in "Clements Harmful Vision" (figure A6), the flag appears for about half a second in a small frame, and in "Ayotte Liberal" (figure A7), it appears for about one second as part of a newspaper masthead. In the next two, a flag pattern appears on an article of clothing—in "Reid Garland Welch" (figure A8), a construction worker's hardhat features a red-and-white striped flag pattern, and in "Toomey Generations (revised)" (figure

A9), the candidate is depicted throughout the ad wearing what appears to be an American flag patterned necktie. Finally, in "Dr. No" (figure A10), flags, in black and white, appear in the background for several seconds.

A9. Variation in coding interface

We made a few changes to the coding interface as the project proceeded. These include dropping several items: (1) We initially coded for sociotropic vs. pocketbook financial appeals; after observing very low reliability we dropped this in favor of simply coding for the presence or absence of economic appeals. We discuss these items further in the conclusion of the paper. (2) We initially coded only fear and enthusiasm; we then added anger and disgust. (3) We added coding for the American flag. (4) We initially coded for mention of a candidate's physical appearance (e.g., hair, makeup, clothing, etc.), and for any specific mention of gender (e.g., "as a woman, I'm running for Congress") or gender-specific role (e.g., "as a father, . . ."). We dropped these items when it became clear that appearance references never appeared, and gender/gender-specific roles were extremely rare. (5) We modified the categories for coding whether each candidate appears in and ad. Initially, the categories were "NO reference to candidate"; "Voice/picture in 'paid for' only"; "Verbal or text reference in ad"; and "Candidate pictured in ad." In later coding this was changed for the favored candidate to "NO reference"; "In 'paid for' only"; "Actual name"; "Picture, video, or audio" and for the opponent, "NO reference"; "In 'paid for' only"; "Vague/generic only" (i.e., "my opponent"); "Picture, video, or audio." Finally, we simplified this for both candidates to "Yes, some reference or information" and "NO reference at all." In all analysis we collapse the more detailed ratings to this final binary. (6) Finally, we reformatted the instructions box to include headings for the different sections ("Economics," "Emotional appeals," etc.) Except for the economic coding, we found no evidence that these changes affected reliability or validity of any coding.

A10. Compensation of mTurk workers and Research Assistants

Initially we paid mTurk workers $0.07 per ad for the first wave of 1,231 ads. Based on feedback from coders, our analysis of the time workers spent, and the addition of a few coding items, we increased the rate to $0.11 per ad for the bulk of the coding (2,659 ads). In the final round of 617 ads we reduced the rate to $0.10, which we found sufficient to attract and retain a large number of workers. Including the 20 percent Amazon.com commission, this works out to $0.12 per ad for a single coding, or $0.60 to have each ad coded five times.

We paid our research assistants the standard university rate of $11 per hour, plus 6% fringe. This worked out to $0.56 per ad coding. In a full-scale project the costs for research assistants would vary. On the one hand, training time would be amortized across more ads, which would lower the per-ad costs. On the other hand, we would double-code a subset of ads to allow reliability analysis. Ultimately, research assistants would likely cost between $0.50 and $0.60 per ad. Thus, the two workforces have comparable cost assuming we have each ad coded by four or five online workers.

Appendix References

Adda, Gilles, Joseph J. Mariani, Laurent Besacier, and Hadrien Gelas. 2013. "Economic and Ethical Background of Crowdsourcing for Speech." In *Crowdsourcing for Speech Processing*: John Wiley & Sons, Ltd, 303-34.

Busarovs, Aleksejs. 2013. "Ethical Aspects of Crowdsourceing, or Is It a Modern Form of Exploitation?" *International Journal of Economics & Business Administration* 1 (1):3-14.

Conger, Anthony J. 1980. "Integration and Generalization of Kappas for Multiple Raters." *Psychological Bulletin* 88 (2):322-8.

Fort, Karën, Gilles Adda, and K. Bretonnel Cohen. 2011. "Amazon Mechanical Turk: Gold Mine or Coal Mine?" *Computational Linguistics* 37 (2):413-20.

Gwet, Kilem L. 2014. *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement among Multiple Raters*. Gaithersburg, MD: Advanced Analytics, LLC.

Hauser, David J., and Norbert Schwarz. 2015. "Attentive Turkers: Mturk Participants Perform Better on Online Attention Checks Than Do Subject Pool Participants." *Behavior Research Methods* 48:400-7.

Kinder, Donald R., and D. Roderick Kiewiet. 1979. "Economic Discontent and Political Behavior: The Role of Personal Grievances and Collective Economic Judgments in Congressional Voting." *American Journal of Political Science* 23 (3):495-527.

———. 1981. "Sociotropic Politics: The American Case." *British Journal of Political Science* 11 (2):129-61.

Klein, Daniel. 2017. Kappaetc: Stata Module to Evaluate Interrater Agreement. Software. Version 1.1.0. https://ideas.repec.org/c/boc/bocode/s458283.html.

Marinova, Dani M. 2016. "On the Use of Crowdsourcing Labor Markets in Research." *Perspectives on Politics* 14 (2):422-31.

Meyer, Doug. 2016. "The Gentle Neoliberalism of Modern Anti-Bullying Texts: Surveillance, Intervention, and Bystanders in Contemporary Bullying Discourse." *Sexuality Research and Social Policy* 13 (4):356-70.

Williamson, Vanessa. 2016. "On the Ethics of Crowdsourced Research." *PS: Political Science & Politics* 49 (1):77-81.

Table A1: Inter-coder reliability by item (Krippendorff's α)

	Research assistants	mTurk workers	mTurk vs. RA	mTurk vs. RA (%)
Flag appears	**0.71**	**0.52**	**−0.19**	**−26%**
FC appears or mentioned	0.93	0.80	−0.12	−13%
OC appears or mentioned	0.87	0.87	−0.00	−0%
Average for candidate appears	**0.90**	**0.83**	**−0.06**	**−7%**
Economic appeal	**0.54**	**0.37**	**−0.17**	**−32%**
Optimistic economic	0.65	0.58	−0.07	−11%
Pessimistic economic	0.72	0.58	−0.14	−20%
Average for Economic tone	**0.68**	**0.58**	**−0.10**	**−15%**
Emotion: enthusiasm	0.29	0.39	+0.10	+34%
Emotion: fear	0.26	0.33	+0.07	+25%
Emotion: anger	0.48	0.38	−0.10	−21%
Emotion: disgust	0.20	0.33	+0.13	+65%
Average for emotions	**0.31**	**0.36**	**+0.05**	**+16%**
Emotion: enthusiasm (dichot)	0.32	0.40	+0.07	+23%
Emotion: fear (dichot)	0.22	0.33	+0.10	+46%
Emotion: anger (dichot)	0.50	0.41	−0.09	−19%
Emotion: disgust (dichot)	0.21	0.35	+0.14	+69%
Average for dichotomized emotions[1]	**0.31**	**0.37**	**+0.06**	**+18%**
FC competence	0.45	0.36	−0.09	−20%
FC strong leader	0.35	0.40	+0.05	+13%
FC integrity	0.43	0.29	−0.14	−32%
FC empathy	0.36	0.29	−0.07	−20%
Average for FC traits	**0.40**	**0.34**	**−0.06**	**−16%**
OC incompetence	0.30	0.31	+0.02	+5%
OC weak leader	0.41	0.23	−0.19	−45%
OC lacks integrity	0.61	0.43	−0.19	−30%
OC cold	0.33	0.28	−0.05	−16%
Average for OC traits	**0.41**	**0.31**	**−0.10**	**−25%**
FC ideology	0.63	0.40	−0.23	−36%
OC ideology	0.64	0.41	−0.22	−35%
Average for ideology	**0.63**	**0.41**	**−0.23**	**−36%**

Entries are Krippendorff's α for multiple raters; with ordinal weights for three-point emotion items and quadratic weights for 101-point ideology items, averaged across individual items. Coefficients calculated by Stata add-on `kappaetc` (Klein 2017).

[1] Emotion coding is on a three-point scale (strong, weak, none); dichotomized versions collapse strong and weak.

Rows in boldface correspond to those in the summary tables in the main paper.

Table A2: Reliability gains from aggregation (Krippendorff's α)

	mTurk workers (on meta subset)	mTurk meta-coders	Difference: meta-coder gain	Research assistants (on meta subset)	meta-mTurk vs. RA	meta-mTurk vs. RA (%)
FC appears or mentioned	0.84	0.97	+0.13	0.96	+0.01	+1%
OC appears or mentioned	0.88	0.96	+0.08	0.88	+0.08	+9%
Average for candidate appears	**0.86**	**0.97**	**+0.10**	**0.92**	**+0.05**	**+5%**
Economic appeal	**0.34**	**0.39**	**+0.06**	**0.53**	**−0.14**	**−26%**
Optimistic economic	0.61	0.71	+0.10	0.65	+0.07	+10%
Pessimistic economic	0.60	0.71	+0.11	0.68	+0.02	+4%
Average for Economic tone	**0.60**	**0.71**	**+0.10**	**0.66**	**+0.05**	**+7%**
Emotion: enthusiasm	0.25	0.37	+0.11	0.30	+0.07	+24%
Emotion: fear	0.38	0.63	+0.24	0.24	+0.39	+160%
Emotion: anger	0.30	0.41	+0.12	0.49	−0.07	−15%
Emotion: disgust	0.22	0.47	+0.24	0.10	+0.37	+371%
Average for emotions	**0.29**	**0.47**	**+0.18**	**0.28**	**+0.19**	**+67%**
Emotion: enthusiasm (dichot)	0.27	0.48	+0.21	0.33	+0.15	+45%
Emotion: fear (dichot)	0.39	0.65	+0.26	0.17	+0.48	+294%
Emotion: anger (dichot)	0.33	0.52	+0.20	0.48	+0.04	+8%
Emotion: disgust (dichot)	0.24	0.35	+0.11	0.09	+0.26	+274%
Average for dichotomized emotions[1]	**0.31**	**0.50**	**+0.19**	**0.27**	**+0.23**	**+87%**
FC competence	0.36	0.63	+0.28	0.48	+0.16	+33%
FC strong leader	0.42	0.72	+0.30	0.36	+0.36	+99%
FC integrity	0.26	0.41	+0.15	0.44	−0.03	−7%
FC empathy	0.25	0.47	+0.22	0.34	+0.13	+37%
Average for FC traits	**0.32**	**0.56**	**+0.24**	**0.41**	**+0.15**	**+38%**
OC incompetence	0.35	0.50	+0.15	0.24	+0.26	+109%
OC weak leader	0.26	0.42	+0.16	0.42	−0.00	−0%
OC lacks integrity	0.41	0.55	+0.14	0.58	−0.03	−5%
OC cold	0.31	0.47	+0.16	0.26	+0.21	+81%
Average for OC traits	**0.33**	**0.49**	**+0.15**	**0.38**	**+0.11**	**+29%**
FC ideology	0.41	0.66	+0.26	0.58	+0.09	+15%
OC ideology	0.46	0.69	+0.23	0.75	−0.07	−9%
Average for ideology	**0.43**	**0.68**	**+0.24**	**0.66**	**+0.01**	**+2%**

Entries are Krippendorff's α for multiple raters; with ordinal weights for three-point emotion items and quadratic weights for 101-point ideology items, averaged across individual items. Coefficients calculated by Stata add-on `kappaetc` (Klein 2017).

Meta-coders are created by averaging five randomly-selected mTurk coders, and then rounding the result to generate a categorical code. Analysis restricted to ads for which we have more than one meta-coder.

[1] Emotion coding is on a three-point scale (strong, weak, none); dichotomized versions collapse strong and weak.

Rows in boldface correspond to those in the summary tables in the main paper.

Table A3: Inter-coder reliability by item (RMSE)

	Research assistants	mTurk workers	mTurk vs. RA	mTurk vs. RA (%)
Flag appears	**0.34**	**0.33**	**−0.008**	**−2%**
FC appears or mentioned	0.14	0.18	+0.034	+24%
OC appears or mentioned	0.23	0.19	−0.040	−18%
Average for candidate appears	**0.18**	**0.18**	**−0.003**	**−2%**
Economic appeal	**0.43**	**0.44**	**+0.015**	**+4%**
Optimistic economic	0.37	0.35	−0.018	−5%
Pessimistic economic	0.30	0.34	+0.047	+16%
Average for Economic tone	**0.33**	**0.35**	**+0.015**	**+4%**
Emotion: enthusiasm	0.40	0.36	−0.042	−11%
Emotion: fear	0.36	0.37	+0.013	+4%
Emotion: anger	0.39	0.36	−0.035	−9%
Emotion: disgust	0.36	0.37	+0.011	+3%
Average for emotions	**0.38**	**0.36**	**−0.013**	**−4%**
Emotion: enthusiasm (dichot)	0.50	0.43	−0.077	−15%
Emotion: fear (dichot)	0.49	0.45	−0.041	−8%
Emotion: anger (dichot)	0.45	0.43	−0.019	−4%
Emotion: disgust (dichot)	0.48	0.45	−0.030	−6%
Average for dichotomized emotions[1]	**0.48**	**0.44**	**−0.042**	**−9%**
FC competence	0.46	0.44	−0.024	−5%
FC strong leader	0.51	0.43	−0.086	−17%
FC integrity	0.37	0.45	+0.080	+21%
FC empathy	0.43	0.41	−0.014	−3%
Average for FC traits	**0.44**	**0.43**	**−0.011**	**−3%**
OC incompetence	0.43	0.45	+0.019	+4%
OC weak leader	0.42	0.44	+0.021	+5%
OC lacks integrity	0.37	0.41	+0.042	+11%
OC cold	0.44	0.42	−0.026	−6%
Average for OC traits	**0.42**	**0.43**	**+0.014**	**+3%**
FC ideology	0.16	0.21	+0.052	+32%
OC ideology	0.18	0.23	+0.058	+33%
Average for ideology	**0.17**	**0.22**	**+0.055**	**+33%**

Entries are root mean squared error among coding decision, calculated as described in the text.

Table A4: Reliability gains due to meta-coders by item (RMSE)

	mTurk workers (on meta subset)	mTurk meta-coders	Difference: meta-coder gain	Research assistants (on meta subset)	meta-mTurk vs. RA	meta-mTurk vs. RA (%)
FC appears or mentioned	0.17	0.09	−0.076	0.09	−0.002	−2%
OC appears or mentioned	0.17	0.08	−0.092	0.18	−0.099	−56%
Average for candidate appears	**0.17**	**0.09**	−0.084	**0.14**	**−0.050**	**−37%**
Economic appeal	**0.44**	**.**	—	**0.38**	—	—
Optimistic economic	0.30	0.19	−0.112	0.33	−0.136	−42%
Pessimistic economic	0.30	0.20	−0.103	0.28	−0.077	−28%
Average for Economic tone	**0.30**	**0.20**	−0.107	**0.30**	**−0.106**	**−35%**
Emotion: enthusiasm	0.37	0.20	−0.168	0.34	−0.141	−41%
Emotion: fear	0.33	0.16	−0.174	0.31	−0.151	−49%
Emotion: anger	0.37	.	—	0.34	—	—
Emotion: disgust	0.37	.	—	0.31	—	—
Average for emotions	**0.36**	**0.18**	−0.178	**0.33**	**−0.146**	**−45%**
Emotion: enthusiasm (dichot)	0.44	0.24	−0.198	0.43	−0.190	−44%
Emotion: fear (dichot)	0.40	0.19	−0.212	0.47	−0.274	−59%
Emotion: anger (dichot)	0.45	.	—	0.40	—	—
Emotion: disgust (dichot)	0.48	.	—	0.44	—	—
Average for dichotomized emotions[1]	**0.44**	**0.22**	−0.225	**0.43**	**−0.217**	**−50%**
FC competence	0.42	0.20	−0.218	0.39	−0.189	−48%
FC strong leader	0.40	0.19	−0.212	0.44	−0.256	−58%
FC integrity	0.43	0.21	−0.211	0.29	−0.077	−26%
FC empathy	0.40	0.17	−0.224	0.37	−0.192	−52%
Average for FC traits	**0.41**	**0.19**	−0.216	**0.37**	**−0.178**	**−48%**
OC incompetence	0.41	0.19	−0.221	0.37	−0.183	−49%
OC weak leader	0.42	0.19	−0.228	0.40	−0.208	−52%
OC lacks integrity	0.39	0.19	−0.201	0.34	−0.153	−44%
OC cold	0.39	0.17	−0.223	0.41	−0.235	−58%
Average for OC traits	**0.40**	**0.19**	−0.218	**0.38**	**−0.195**	**−51%**
FC ideology	0.20	0.13	−0.072	0.15	−0.018	−12%
OC ideology	0.22	0.13	−0.095	0.13	−0.002	−1%
Average for ideology	**0.21**	**0.13**	−0.083	**0.14**	**−0.010**	**−7%**

Entries are root mean squared error among coding decision, calculated as described in the text.

Table A5: Inter-coder reliability statistics by item (Conger's κ)

	Research assistants	mTurk workers	mTurk vs. RA	mTurk vs. RA (%)
Flag appears	**0.69**	**0.61**	**−0.08**	**−11%**
FC appears or mentioned	0.92	0.75	−0.17	−18%
OC appears or mentioned	0.87	0.85	−0.01	−1%
Average for candidate appears	**0.89**	**0.80**	**−0.09**	**−10%**
Economic appeal	**0.55**	**0.37**	**−0.17**	**−31%**
Optimistic economic	0.63	0.57	−0.06	−10%
Pessimistic economic	0.73	0.59	−0.14	−19%
Average for Economic tone	**0.68**	**0.58**	**−0.10**	**−15%**
Emotion: enthusiasm	0.29	0.41	+0.12	+40%
Emotion: fear	0.28	0.34	+0.06	+23%
Emotion: anger	0.47	0.41	−0.06	−12%
Emotion: disgust	0.22	0.39	+0.16	+73%
Average for emotions	**0.31**	**0.39**	**+0.07**	**+23%**
Emotion: enthusiasm (dichot)	0.32	0.41	+0.08	+26%
Emotion: fear (dichot)	0.27	0.29	+0.02	+8%
Emotion: anger (dichot)	0.49	0.37	−0.12	−25%
Emotion: disgust (dichot)	0.25	0.35	+0.09	+37%
Average for dichotomized emotions[1]	**0.33**	**0.35**	**+0.02**	**+6%**
FC competence	0.43	0.36	−0.08	−18%
FC strong leader	0.35	0.39	+0.05	+13%
FC integrity	0.38	0.30	−0.08	−20%
FC empathy	0.37	0.23	−0.14	−38%
Average for FC traits	**0.38**	**0.32**	**−0.06**	**−16%**
OC incompetence	0.29	0.29	+0.01	+2%
OC weak leader	0.44	0.16	−0.27	−62%
OC lacks integrity	0.60	0.41	−0.20	−33%
OC cold	0.36	0.18	−0.18	−49%
Average for OC traits	**0.42**	**0.26**	**−0.16**	**−38%**
FC ideology	0.65	0.45	−0.20	−31%
OC ideology	0.58	0.45	−0.13	−23%
Average for ideology	**0.62**	**0.45**	**−0.17**	**−27%**

Entries are Conger's κ for multiple raters; with ordinal weights for three-point emotion items and quadratic weights for 101-point ideology items, averaged across individual items. Coefficients calculated by Stata add-on `kappaetc` (Klein 2017).

[1] Emotion coding is on a three-point scale (strong, weak, none); dichotomized versions collapse strong and weak.

Rows in boldface correspond to those in the summary tables in the main paper.

Table A6: Reliability gains from aggregation, by item (Conger's κ)

	mTurk workers (on meta subset)	mTurk meta-coders	Difference: meta-coder gain	Research assistants (on meta subset)	meta-mTurk vs. RA	meta-mTurk vs. RA (%)
FC appears or mentioned	0.83	0.98	+0.15	0.97	+0.01	+1%
OC appears or mentioned	0.88	0.95	+0.07	0.90	+0.05	+6%
Average for candidate appears	**0.85**	**0.96**	**+0.11**	**0.93**	**+0.03**	**+4%**
Economic appeal	**0.34**	**0.41**	**+0.07**	**0.55**	**−0.13**	**−25%**
Optimistic economic	0.59	0.70	+0.11	0.61	+0.09	+15%
Pessimistic economic	0.59	0.70	+0.11	0.66	+0.04	+6%
Average for Economic tone	**0.59**	**0.70**	**+0.11**	**0.63**	**+0.07**	**+10%**
Emotion: enthusiasm	0.26	0.36	+0.09	0.30	+0.06	+20%
Emotion: fear	0.38	0.63	+0.25	0.26	+0.37	+144%
Emotion: anger	0.34	0.36	+0.02	0.44	−0.08	−19%
Emotion: disgust	0.31	0.46	+0.15	0.14	+0.32	+228%
Average for emotions	**0.32**	**0.45**	**+0.13**	**0.28**	**+0.17**	**+59%**
Emotion: enthusiasm (dichot)	0.30	0.51	+0.21	0.33	+0.18	+55%
Emotion: fear (dichot)	0.37	0.66	+0.28	0.16	+0.50	+319%
Emotion: anger (dichot)	0.35	0.53	+0.18	0.44	+0.09	+21%
Emotion: disgust (dichot)	0.28	0.34	+0.06	0.11	+0.23	+213%
Average for dichotomized emotions[1]	**0.32**	**0.51**	**+0.18**	**0.26**	**+0.25**	**+97%**
FC competence	0.34	0.61	+0.27	0.45	+0.16	+35%
FC strong leader	0.40	0.70	+0.30	0.37	+0.33	+91%
FC integrity	0.28	0.41	+0.13	0.43	−0.02	−5%
FC empathy	0.23	0.56	+0.33	0.37	+0.19	+52%
Average for FC traits	**0.31**	**0.57**	**+0.26**	**0.40**	**+0.16**	**+41%**
OC incompetence	0.32	0.47	+0.15	0.27	+0.20	+72%
OC weak leader	0.18	0.47	+0.29	0.42	+0.05	+11%
OC lacks integrity	0.40	0.57	+0.16	0.55	+0.02	+4%
OC cold	0.25	0.49	+0.24	0.29	+0.20	+69%
Average for OC traits	**0.29**	**0.50**	**+0.21**	**0.38**	**+0.12**	**+31%**
FC ideology	0.39	0.68	+0.29	0.60	+0.07	+12%
OC ideology	0.43	0.65	+0.22	0.74	−0.09	−12%
Average for ideology	**0.41**	**0.66**	**+0.25**	**0.67**	**−0.01**	**−1%**

Entries are Conger's κ for multiple raters; with ordinal weights for three-point emotion items and quadratic weights for 101-point ideology items, averaged across individual items. Coefficients calculated by Stata add-on `kappaetc` (Klein 2017).

Meta-coders are created by averaging five randomly-selected mTurk coders, and then rounding the result to generate a categorical code. Analysis restricted to ads for which we have more than one meta-coder.

[1] Emotion coding is on a three-point scale (strong, weak, none); dichotomized versions collapse strong and weak.

Rows in boldface correspond to those in the summary tables in the main paper.

Table A7: Gains to aggregation for meta-coders made up of between 2 and 8 individuals

Size of meta-coder	Average RMSE relative to single coder	Incremental reduction
1	1.000	
2	0.798	0.202
3	0.685	0.113
4	0.616	0.069
5	0.574	0.042
6	0.542	0.032
7	0.518	0.024
8	0.499	0.020

Table depicts RMSE for meta-coders, relative to a single coder.

Table A8: Impact of coder characteristics on coding time

	Time to code ad (seconds)
Coder's education: Some college	−8.161
	(6.573)
Coder's education: Associate degree	−0.036
	(6.889)
Coder's education: Bachelor degree	−0.440
	(5.704)
Coder's education: Graduate degree	7.114
	(7.005)
Coder's political knowledge	3.409
	(60.547)
Coder's political knowledge × Coder's political knowledge	−0.633
	(42.683)
Female coder	5.743^
	(3.170)
Coder is student	13.429*
	(6.432)
Republican coder	−15.149**
	(4.566)
Democratic coder	−16.838**
	(4.498)
Coder age 26-30	1.851
	(5.824)
Coder age 31-35	10.902^
	(5.819)
Coder age 36-40	14.090*
	(6.584)
Coder age 41-50	26.050**
	(6.806)
Coder age 51+	32.619**
	(6.384)
Coder's income: 20k-40k	−11.041^
	(6.306)
Coder's income: 40k-80k	−9.582
	(6.177)
Coder's income: 80k+	−18.043*
	(7.091)
African American Coder	11.504
	(7.991)
Asian American Coder	−5.003
	(4.882)
Latinx Coder	14.648
	(11.312)
Yes, occasionally	−5.269
	(4.503)
Yes, rarely	−12.128*
	(5.209)
No, never	−10.677*
	(4.522)
Coded 11-49 ads overall	−34.755**
	(5.658)
Coded 50-99 ads overall	−41.892**
	(6.306)
Coded 100+ ads overall	−47.299**
	(5.539)
Ads 11-49	−24.223**
	(2.197)
Ads 50-99	−30.186**
	(2.953)
Ads 100-199	−34.199**
	(3.446)
Ads 200+	−36.493**
	(4.059)
Intercept	150.699**
	(24.566)

N	26,013
Std. error of regression	46.34
R^2	0.21

Table A9: Impact of ad characteristics on coding time

	Time to code ad (seconds)
Race type: Senate race	0.099
	(3.038)
Ad sponsor: party	−4.973**
	(1.094)
Ad sponsor: coordinated between candidate and party	−1.256
	(1.637)
Ad sponsor: interest group or other	−4.395**
	(0.994)
Ad tone: promote	−6.873**
	(1.119)
Ad tone: attack	−1.674
	(1.205)
Ad focus: neither	3.717
	(2.518)
Ad focus: personal characteristics	0.838
	(1.108)
Ad focus: both personal characteristics and policy matters	0.391
	(0.656)
Intercept	76.708**
	(2.997)
N	25,994
Std. error of regression	52.06
R^2	0.00

Run among mTurk worker coding of 30-second ads. Robust standard errors, clustered by worker.
** p<0.01; * p<0.05; ^ p<0.10 two tailed.

Table A10: Impact of time spent coding on reliability

	Absolute decision-level "error"
Average for candidate appears	−0.107** (0.011)
Economic appeal	0.168** (0.018)
Average for economic tone	0.023* (0.012)
Average for emotions	0.100** (0.012)
Average for FC traits	0.151** (0.014)
Average for OC traits	0.134** (0.015)
Average for ideology	0.016 (0.011)
Time 40-54 (second quartile)	0.020 (0.016)
Time 55-85 (third quartile)	0.023^ (0.012)
Time 86-132 (75th-90th pctile)	0.008 (0.013)
Time >134 (90th-100th pctile)	0.008 (0.018)
Average for candidate appears × Time 40-54 (second quartile)	−0.015 (0.017)
Average for candidate appears × Time 55-85 (third quartile)	−0.019 (0.013)
Average for candidate appears × Time 86-132 (75th-90th pctile)	−0.000 (0.015)
Average for candidate appears × Time >134 (90th-100th pctile)	0.006 (0.018)
Economic appeal × Time 40-54 (second quartile)	−0.023 (0.022)
Economic appeal × Time 55-85 (third quartile)	−0.032 (0.023)
Economic appeal × Time 86-132 (75th-90th pctile)	−0.035^ (0.020)
Economic appeal × Time >134 (90th-100th pctile)	−0.021 (0.027)
Average for economic tone × Time 40-54 (second quartile)	−0.026 (0.017)
Average for economic tone × Time 55-85 (third quartile)	−0.015 (0.015)
Average for economic tone × Time 86-132 (75th-90th pctile)	0.011 (0.016)
Average for economic tone × Time >134 (90th-100th pctile)	0.005 (0.021)
Average for emotions × Time 40-54 (second quartile)	−0.005 (0.017)
Average for emotions × Time 55-85 (third quartile)	−0.005 (0.013)
Average for emotions × Time 86-132 (75th-90th pctile)	0.018 (0.016)
Average for emotions × Time >134 (90th-100th pctile)	0.014 (0.019)
Average for FC traits × Time 40-54 (second quartile)	−0.021 (0.018)
Average for FC traits × Time 55-85 (third quartile)	−0.016 (0.015)
Average for FC traits × Time 86-132 (75th-90th pctile)	−0.002 (0.019)
Average for FC traits × Time >134 (90th-100th pctile)	−0.006 (0.020)
Average for OC traits × Time 40-54 (second quartile)	0.008

	(0.018)
Average for OC traits × Time 55-85 (third quartile)	0.004
	(0.016)
Average for OC traits × Time 86-132 (75th-90th pctile)	0.027
	(0.018)
Average for OC traits × Time >134 (90th-100th pctile)	0.023
	(0.020)
Average for ideology × Time 40-54 (second quartile)	−0.016
	(0.016)
Average for ideology × Time 55-85 (third quartile)	−0.021
	(0.013)
Average for ideology × Time 86-132 (75th-90th pctile)	−0.003
	(0.015)
Average for ideology × Time >134 (90th-100th pctile)	0.002
	(0.018)
Intercept	0.154**
	(0.009)
N	388,509
Std. error of regression	0.27
R^2	0.09

Table A11: Impact of time spent coding on validity

	FC ideology
Favored candidate DW-NOMINATE	0.279**
	(0.032)
Time 40-54 (second quartile)	−0.008
	(0.010)
Time 55-85 (third quartile)	−0.006
	(0.011)
Time 86-132 (75th-90th pctile)	−0.025^
	(0.013)
Time >134 (90th-100th pctile)	−0.011
	(0.013)
Time 40-54 (second quartile) × Favored candidate DW-NOMINATE	0.009
	(0.027)
Time 55-85 (third quartile) × Favored candidate DW-NOMINATE	0.031
	(0.033)
Time 86-132 (75th-90th pctile) × Favored candidate DW-NOMINATE	0.056
	(0.037)
Time >134 (90th-100th pctile) × Favored candidate DW-NOMINATE	0.067^
	(0.039)
Intercept	0.570**
	(0.011)
N	8,923
Std. error of regression	0.21
R^2	0.25

Table A12: Coder learning or fatigue (reliability)

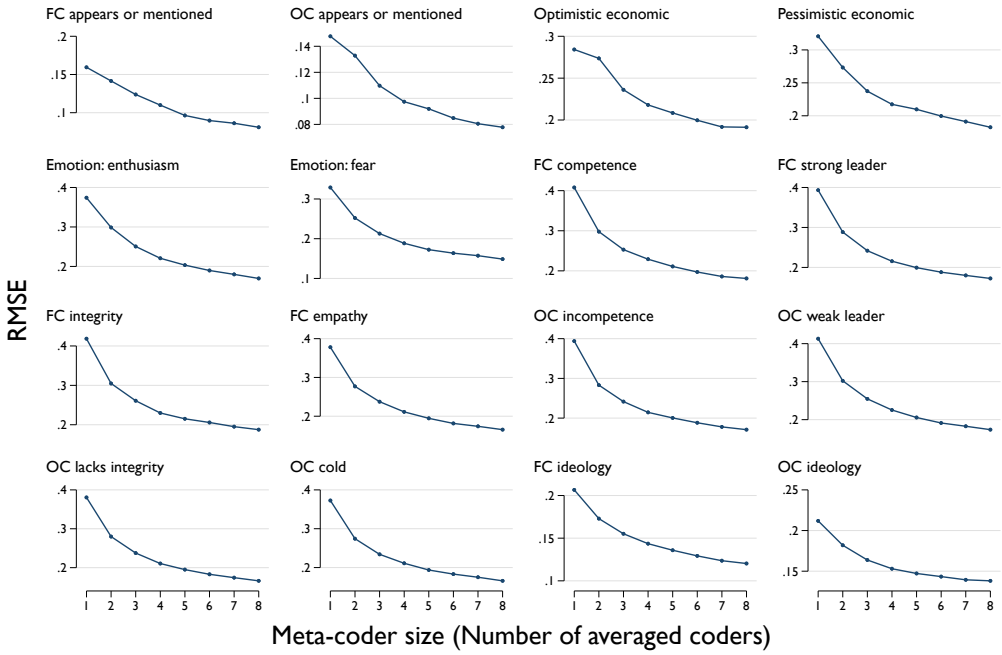
	Absolute decision-level "error"
Average for candidate appears	−0.123** (0.013)
Economic appeal	0.104** (0.024)
Average for economic tone	0.021 (0.016)
Average for emotions	0.088** (0.014)
Average for FC traits	0.106** (0.014)
Average for OC traits	0.104** (0.014)
Average for ideology	0.006 (0.013)
Ads 11-49	−0.021 (0.020)
Ads 50-99	−0.008 (0.019)
Ads 100-199	−0.007 (0.019)
Ads 200+	−0.020* (0.010)
Average for candidate appears × Ads 11-49	0.011 (0.020)
Average for candidate appears × Ads 50-99	−0.008 (0.019)
Average for candidate appears × Ads 100-199	−0.005 (0.019)
Average for candidate appears × Ads 200+	0.013 (0.011)
Economic appeal × Ads 11-49	0.025 (0.031)
Economic appeal × Ads 50-99	0.030 (0.030)
Economic appeal × Ads 100-199	0.042 (0.033)
Economic appeal × Ads 200+	0.051* (0.024)
Average for economic tone × Ads 11-49	0.001 (0.022)
Average for economic tone × Ads 50-99	−0.016 (0.022)
Average for economic tone × Ads 100-199	−0.020 (0.023)
Average for economic tone × Ads 200+	−0.012 (0.017)
Average for emotions × Ads 11-49	0.023 (0.020)
Average for emotions × Ads 50-99	0.000 (0.020)
Average for emotions × Ads 100-199	0.004 (0.022)
Average for emotions × Ads 200+	0.014 (0.017)
Average for FC traits × Ads 11-49	0.020 (0.020)
Average for FC traits × Ads 50-99	0.020 (0.020)
Average for FC traits × Ads 100-199	0.029 (0.018)
Average for FC traits × Ads 200+	0.058** (0.011)
Average for OC traits × Ads 11-49	0.030

	(0.021)
Average for OC traits × Ads 50-99	0.027
	(0.019)
Average for OC traits × Ads 100-199	0.024
	(0.020)
Average for OC traits × Ads 200+	0.072**
	(0.015)
Average for ideology × Ads 11-49	0.003
	(0.019)
Average for ideology × Ads 50-99	−0.010
	(0.018)
Average for ideology × Ads 100-199	−0.017
	(0.018)
Intercept	0.183**
	(0.013)
N	376,710
Std. error of regression	0.27
R^2	0.10

Table A13: Coder learning or fatigue (validity)

	FC ideology
Favored candidate DW-NOMINATE	0.308**
	(0.023)
Ads 11-49	0.003
	(0.010)
Ads 50-99	−0.010
	(0.013)
Ads 100-199	−0.006
	(0.013)
Ads 200+	0.012
	(0.017)
Ads 11-49 × Favored candidate DW-NOMINATE	0.007
	(0.024)
Ads 50-99 × Favored candidate DW-NOMINATE	−0.001
	(0.028)
Ads 100-199 × Favored candidate DW-NOMINATE	0.008
	(0.031)
Ads 200+ × Favored candidate DW-NOMINATE	−0.011
	(0.036)
Intercept	0.558**
	(0.009)
N	8,586
Std. error of regression	0.21
R^2	0.25

Figure A1: Decreasing RMSE as a function of meta-coder size



Results averaged across ten replications in which RMSE was calculated for meta-coder made up of 1 through 8 randomly-selected individual coders. Error calculated relative to 'truth' that is calculated from the other coders. Analysis on subset of 221 ads for which we have 18 or more mTurk coders

Table A14: Demographics of mTurk workers and the American public

	mTurk coders	2016 ANES
	%	%
Gender		
Male	48.2	48.0
Female	51.8	52.0
Total	100.0	100.0
Age		
18-24	10.7	12.2
25-34	39.7	16.8
35-44	27.3	15.2
45-54	14.5	17.7
55+	7.8	38.0
Total	100.0	100.0
Coder racial/ethnic identification		
White	78.9	68.3
Black or African-American	6.5	10.8
Asian	6.3	2.7
Hispanic, Latino, or Spanish	2.5	10.7
Other	1.1	1.7
Multiple or mixed race	4.4	5.2
(not specified)	0.4	0.6
Total	100.0	100.0
Education		
Less than HS	0.8	9.1
High school graduate or GED	10.9	28.9
Some college, no degree	27.4	18.7
Associate degree	15.2	12.2
Bachelor's degree	35.8	18.3
Graduate degree	9.9	12.7
Total	100.0	100.0
Student status		
Non-student	88.2	95.9
Student	11.8	4.1
Total	100.0	100.0
Family Income		
<20k	17.0	17.5
20k-40k	25.2	18.6
40k-80k	38.9	28.5
80k+	18.9	35.4
Total	100.0	100.0
Party Identification		
Republican	20.0	28.3
Independent	35.0	36.6
Democrat	45.1	35.1
Total	100.0	100.0
N	526	4,271

ANES estimates are weighted

Figure A2: Impact of time spent coding on reliability

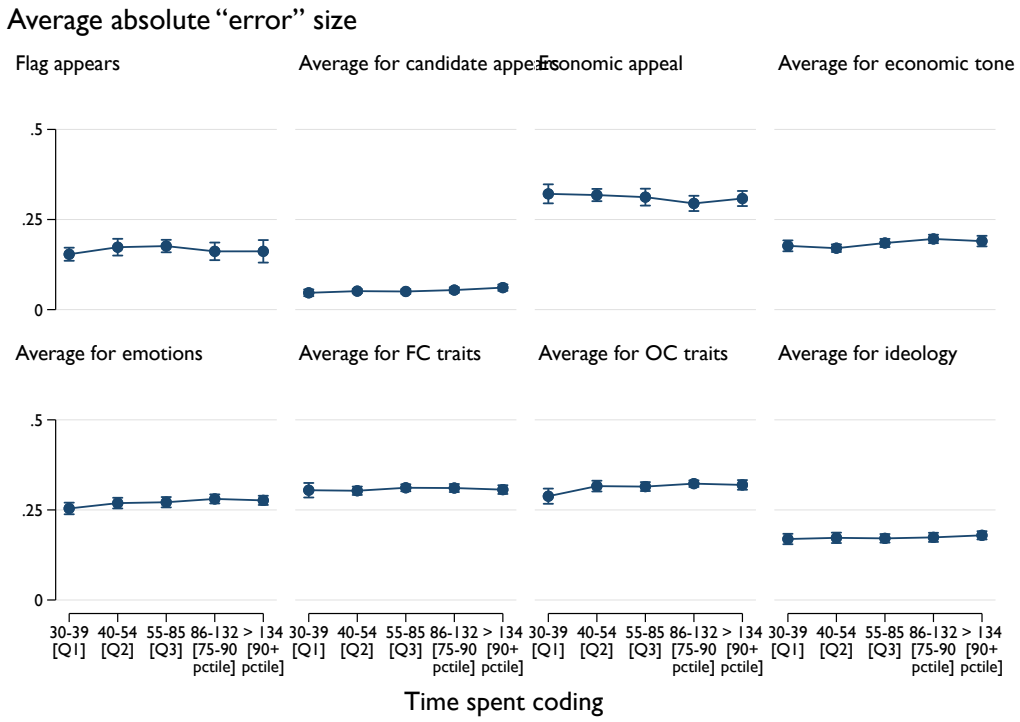


Figure A3: Impact of time spent coding on validity

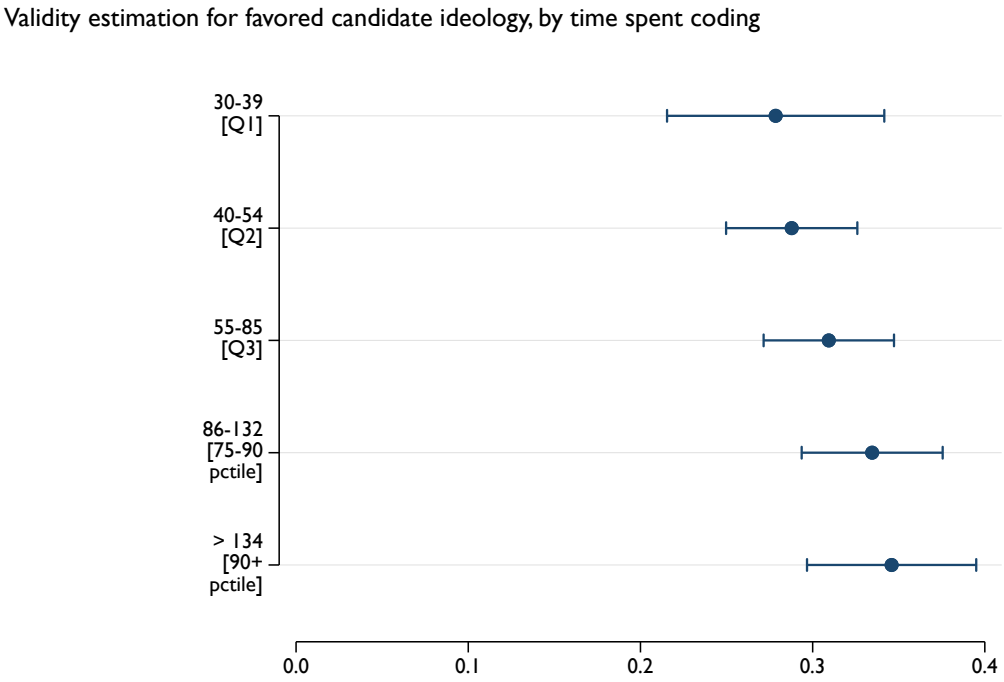


Figure A4: Coder learning or fatigue (reliability)

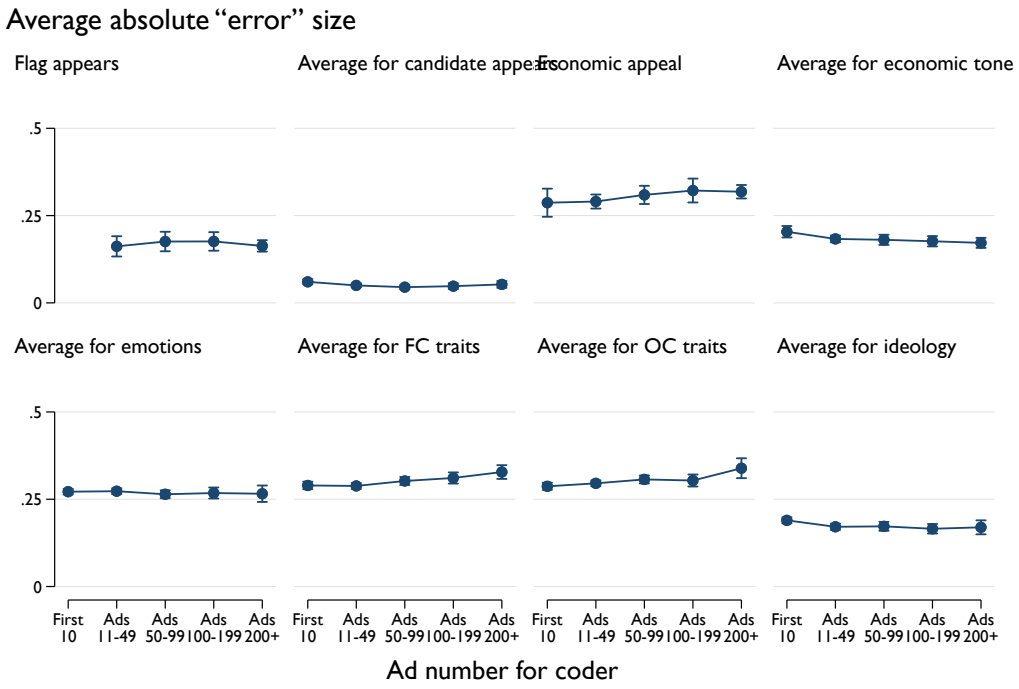


Figure A5: Coder learning or fatigue (validity)

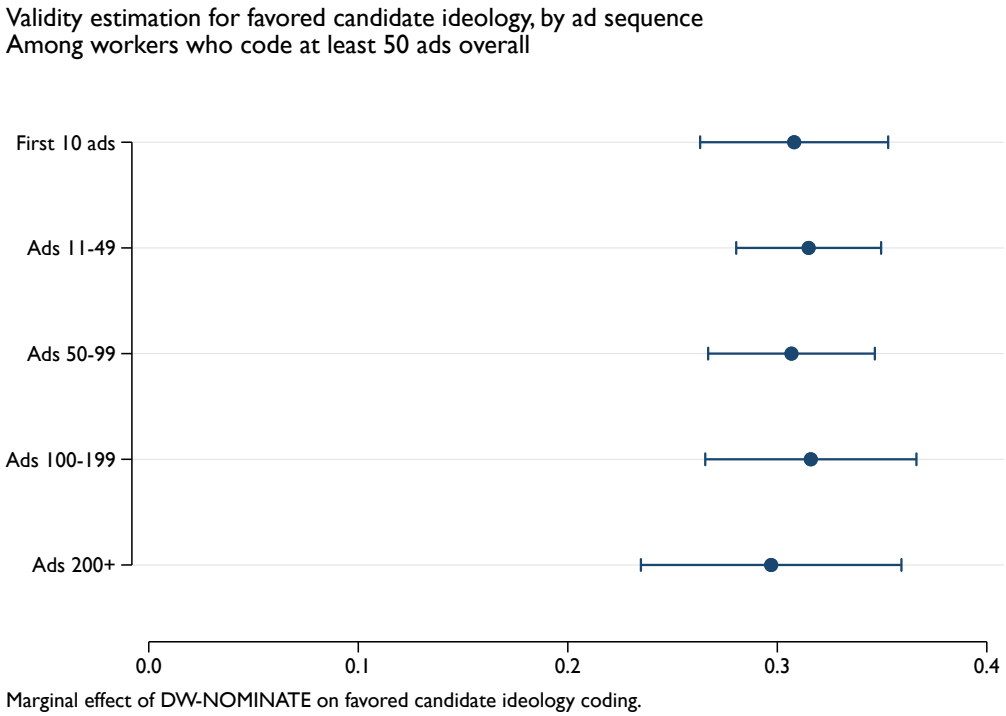


Figure A6: "Clements Harmful Vision" (Tom Clements for US Senate, SC)

Figure A7: "Ayotte Liberal" (Kelly Ayotte for US Senate, NH)

Figure A8: “Reid Garland Welch” (Harry Reid for US Senate, NV)

Figure A9: “Toomey Generations (Revised)” (Pat Toomey for US Senate, PA)

Figure A10: “Dr. No” (Dan Connolly for Congress, PA-08)