Political Science Research and Methods (2019), page 1 of 16
doi:10.1017/psrm.2019.4

ORIGINAL ARTICLE

Online coders, open codebooks: New opportunities for content analysis of political communication

Nicholas J. G. Winter[1]*, Adam G. Hughes[2] and Lynn M. Sanders[1]

[1]Department of Politics, University of Virginia, S185 Gibson Hall, 1540 Jefferson Park Ave, Charlottesville, VA and [2]Pew Research Center, Washington, D.C.
*Corresponding author. Email: nwinter@virginia.edu

(Received 23 February 2018; revised 13 November 2018; accepted 28 November 2018)

Abstract

Analyzing audiovisual communication is challenging because its content is highly symbolic and less rule-governed than verbal material. But audiovisual messages are important to understand: they amplify, enrich, and complicate the meaning of textual information. We describe a fully-reproducible approach to analyzing video content using minimally—but systematically—trained online workers. By aggregating the work of multiple coders, we achieve reliability, validity, and costs that equal those of traditional, intensively trained research assistants, with much greater speed, transparency, and replicability. We argue that measurement strategies relying on the "wisdom of the crowd" provide unique advantages for researchers analyzing complex and intricate audiovisual political content.

Key words: automated content analysis; crowdsourcing; political communication; text and content analysis

We are in the midst of a content-analysis revolution driven by computing and crowd-sourcing. New methodological techniques have helped automate the classification of political texts (Benoit et al. 2009; 2016; Grimmer and Stewart 2013). But existing applications overwhelmingly focus on text-based verbal content, rather than the complex mixture of verbal, non-verbal, and visual signals and imagery in contemporary political communications. Visual and auditory content is frequently downplayed because analyzing it is difficult and time-consuming and because there has been less theorizing on how audiovisual media convey meaning.

In this paper we describe a transparent, efficient, and reproducible approach to analyzing audiovisual political communications. Focusing on political television advertising, we demonstrate that crowd-sourced online coders can measure complex audiovisual content well. We show that crowd-sourcing is particularly suited to this task, where coding categories are not reducible to objective rules, and that this approach is more reproducible than traditional content analysis.

We recruited many coders from Amazon's Mechanical Turk (mTurk) online labor market and trained them systematically, yet less intensively than is typically advised. Through a custom-programmed web platform, we supplied coders with a codebook that they used to train themselves without any feedback. We compare this approach with a traditional method relying on research-assistant coders who we trained intensively and interactively in person. Across a broad range of coding tasks, we show that online workers can code as well as traditional research assistants. While online workers make worse coding decisions individually, their lower cost allows us to rate each piece of content repeatedly. Aggregating these coding decisions produces reliability and validity comparable to or better than that using research assistants, at similar net cost.

Online workers offer several additional advantages. First, they are much faster than research assistants. This makes it easier to develop and execute content analyses using coding schemes

customized to the research question, rather than relying on existing coding. Second, multiple coding of each item generates measures of ambiguity and uncertainty, opening avenues for additional analysis. Third, standardized, hands-off training produces a completely open and fully reproducible data creation process. This allows scholars to build on each other's work by replicating and extending their coding protocols.

Benoit et al. (2016) showed that aggregated non-experts recruited online can equal subject-matter experts at classifying ideology in party manifestos. They also highlighted the advantages of online classification for transparent, reproducible, and adaptable data collection. We extend their work in four ways. First, we compare coding by the crowd not with experts but with student research assistants, who are the typical coders for political communications research. Second, we evaluate coding of real-world political communication. Our coders rated complete political texts—political advertisements—rather than discrete sentences or sound bites. Third, we examine coding not simply of ideology but of a wide range of concepts that figure heavily in audiovisual communication, varying from objective and clear-cut to abstract and symbolic. Finally and most importantly, we focus on the complex medium of video. Audiovisual material is intrinsically difficult to code; therefore, it offers a difficult test of online coders' ability to parse meaning from political communication. Our findings thus speak to the analysis of audiovisual content in political advertising and also in the news, infotainment, and non-political advertising.

The medium and the message

Beginning with viewer reactions to televised imagery of Nixon and Kennedy in the 1960 debates, television has made political communication fundamentally visual. The Johnson campaign's 1964 "Daisy" advertisement "transform(ed) American political advertising" (Mann 2011, 61) with its compelling visual juxtaposition of a mushroom cloud with a girl in a field of daisies. Among many examples, Ronald Reagan's 1984 "Morning in America" and Hillary Clinton's 2008 "3 a.m." both used audio and visual elements to convey powerful emotional messages. Yet analysis of audiovisual political advertising and other forms of communication lags behind work on verbal text. The Enlightenment presumption that "verbal arguments are … the primary conduit of reason" may be partly to blame (Grabe and Bucy 2009, 6). Measuring the content of visual communications is harder, both in theory and in practice, compared with text. This has created a huge gap in our ability to analyze political communication.

"Most Americans receive the bulk of their messages about politics from audiovisual media" (Graber and Smith 2005, 492). And for good reason: visual processing takes place in a brain region specifically adapted for the purpose: the visual cortex. Compared with language processing, visual perception is fast and efficient, engages emotion and cognition simultaneously, and creates stronger memories (Grabe and Bucy 2009, 12–21). People make fast, powerful inferences about traits, emotional states, motivations, and intentions based on facial expressions, gestures, tone of voice, and body language (e.g. Masters and Sullivan 1993). For example, Rosenberg et al. (1986, 123) showed that "a single photograph can have a clear impact on voters' judgments regarding a candidate's congressional demeanor, competence, leadership ability, attractiveness, likeableness, and integrity" and can "exercise a strong and consistent influence on electoral choices."

Pictures and sounds also evoke strong emotions that shape processing of verbal material. Brader (2006) shows, for example, that anxiety cues lead viewers to attend to and remember information conveyed by campaign ads, as in Clinton's 2008 "3 a.m." ad that worked to stir anxiety about Obama's readiness to lead in a dangerous world. Irony and humor can also reinforce an ad's message. In "Tank Ride," the 1988 Bush campaign buttressed the claim that Dukakis was soft on defense with footage of him looking "juvenile and foolish as he takes a 'joy ride' in a tank" (Reynolds and Whitlark 1995, 14). Similarly, in "Windsurfing," the 2004 Bush campaign employed video to forge a link between Kerry's policymaking and personal character (Spielvogel 2005). Visuals can also convey messages that do not appear in the verbal channel

at all. This is especially important for racial messages; for example, a mug-shot of William Horton in 1988's "Weekend Passes" primed white Americans' racial resentment without explicitly mentioning race (Mendelberg 2001). Thus, concludes Benoit (2010, 276), "There can be no question that (visual) elements of texts are fundamentally important: These aspects of texts can reinforce the verbal message (e.g., a candidate declaring his or her patriotism with the American flag in the background), contradict the verbal message (irony or sarcasm indicated by tone of voice), or even send a different message (e.g., subtle cues of racism amid protestations of the importance of equal opportunity)."

People believe and remember what they see more than what they hear or read, and "visual messages override other messages when processed simultaneously" (Schill 2012, 122). This follows from the fact that visuals are *iconographic*—that is, symbols that physically resemble the things they represent, in contrast with the arbitrary signs of spoken or written language (Messaris 1997). Iconography produces powerful psychological effects: cognitive reactions to visual representations of actions mirror the reactions of actual participation. Visual images are perceived as real, which increases persuasion while undermining awareness of persuasive effects. Thus, visuals "operate upon us in a manner which suppresses and conceals their ideological function because they appear to record rather than to transform or signify" (Woollacott 1982, 99; cited in Messaris and Abraham 2001).

For these reasons, scholars have long called for more attention to visual political communication. Graber (1987) concluded that coding only verbal content distorts the meaning of political messages. Almost two decades later, though, Graber and Smith (2005, 492) found little evidence that scholars employed visual coding. And in 2012, Schill argued that "the visual aspects of political communication remain one of the least studied and the least understood areas" (119).

Moreover, visual communication lacks the propositional syntax that governs formal language, making it very hard to code. While there are "relatively precise conventions for indicating spatial or temporal relationships among two or more images," Messaris (1997, x) argues, "visual communication is characterized by a lack of means for identifying other ways in which images might be related to each other." Visual communication operates by informal and less well-understood rules, meaning that it cannot convey precise propositional claims (e.g., this policy is a bad idea because it will lead to outcomes X and Y). Instead, visual messages are ambiguous, carrying multiple messages simultaneously. Even when conveying spoken language, the speaker's tone of voice, gender, and other characteristics can shape message reception powerfully (e.g. Strach et al. 2015).

These challenges to coding visual materials impede the cumulative empirical research needed to support theory development. The power of the crowd promises to address these challenges, advancing both empirical knowledge and theory development by generating a more efficient and replicable measurement of real-life political communications.

Online coding interface, coders, and coding tasks

Coding interface

To allow direct comparisons of coding quality, we created a web-based coding portal used by both research assistants and online coders. The research assistant coders accessed the portal directly through our server, while the online workers saw it embedded within the mTurk worker site (Figure 1). The interface showed coders a single ad, a data entry form, and a summary of the coding instructions. Coders could expand the instructions and also download the codebook.[1] They viewed each randomly selected ad one or more times while coding.[2]

[1]Appendix A1 discusses implementation of the portal.

[2]Online workers averaged 72 seconds coding each 30-second ad. Appendix A6 presents analyses of coding time and its (lack of) impact on reliability and validity.

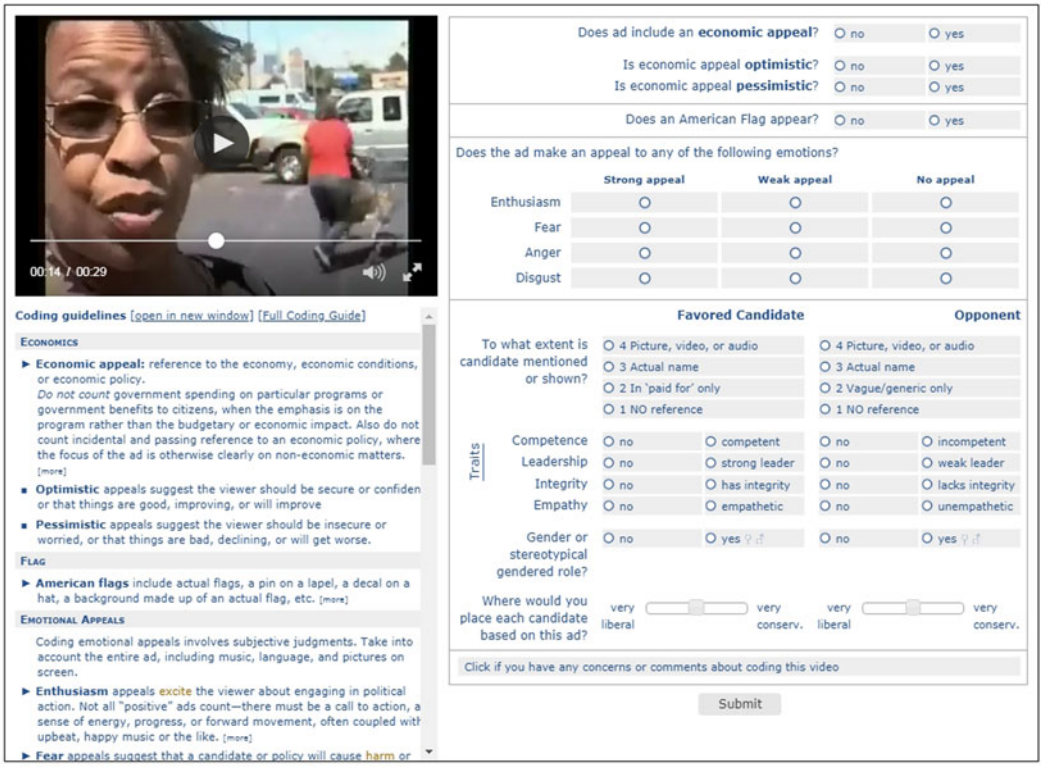

Figure 1. Coding interface

The coding tasks

We coded a diverse array of concepts that (1) range in the likely difficulty of coding; (2) are communicated audiovisually and verbally; (3) vary in the level of political knowledge required to code; and (4) replicate coding provided by the Wesleyan Media Project (WMP).[3]

We drew on Potter and Levine-Donnerstein's (1999, 259) distinction between manifest content that is "on the surface and easily observable," where coding follows simple, objective rules, and latent content, where the meaning underlies the surface. They subdivide latent into *pattern content*, identifiable by coders trained to recognize objectively-defined patterns among symbolic elements, and *projective content*, in which the "elements in the content are symbols that require viewers to access their pre-existing mental schema in order to judge the meaning."

We included two types of manifest content: the presence of the American flag and the mention or appearance in the ad of the favored and opposition candidates. For latent pattern content, we asked coders to identify economic appeals and to classify their tone as optimistic, pessimistic, or both (i.e., mixed). In the latent projective category, we asked coders to identify four types of *emotional appeals*: three provided by WMP (enthusiasm, fear, and anger) plus one additional emotion (disgust); and four *trait attributions* for each ad's favored candidate (competence, strong leadership, integrity, and empathy) and the opponent (incompetence, weak leadership, lack of integrity, and lack of empathy).[4] Finally, we asked coders to assess the *ideological position* of the favored and opposing candidates, as stated or implied by the content of the ad.

[3]Appendix A2 presents additional information about the coding process.

[4]These trait categories followed guidelines developed by Hayes (2011) after scholars who find trait assessment central to candidate evaluation (e.g., Kinder et al. 1980).

Latent projective concepts like these are very hard to code because are complex and symbolic and do not have concise, rule-based definitions. Accordingly, the traditional response is to conduct even more extensive and interactive coder training.[5] We expect, however, that ordinary people are up to the coding task. Although hard to define precisely, most people understand these concepts intuitively, making them something Potter and Levine-Donnerstein (1999, 260) call "primitive concepts." Consider the emotions: most American coders share broad understandings of fear, anger, and other emotions, and we believe that coders can identify and distinguish among them, especially if the instructions cue culturally available cognitive schema. Identifying traits might be harder, though here too we believe that ordinary people will have relatively clear understandings. Coding ideological implication requires more specialized political knowledge, but also relies on coders' ability to recognize novel and creative expressions of ideology that go beyond any set of objective rules for what counts as liberal or conservative.

There is some precedent for using the crowd to code for *textual* latent content. Lind et al. (2017) show that the crowd equals traditional coders in identifying the presence and valence of target evaluations in text sentences. Weber et al. (2018) demonstrate that ordinary citizens can draw on "moral intuitions" to code text. And Budak et al. (2016) show that online workers can classify the partisan slant of newspaper articles.

Content

We analyzed the universe of English-language campaign commercials aired on behalf of 2010 US House and Senate candidates (Fowler et al. 2014). This includes 4,357 unique advertisements (3,016 House and 1,341 Senate) produced by campaigns, political parties, and outside groups on behalf of 662 different candidates. They were aired just over 1.5 million times.

Coders

Research assistants

We hired undergraduate research assistants with recruitment, training, and working conditions that followed standard practice.[6] We emailed advertisements to political science majors and selected six capable students. They completed a background survey and studied the online coding guide; then we trained them in two group meetings. At the first, we explained the codebook, practiced coding several ads together, answered questions, and discussed the guidelines. Between the meetings each research assistant coded a set of 30 practice ads. At the second meeting we discussed and resolved disagreements, further clarified the rules, and answered questions. Then the research assistants began coding; as they did so we stayed in touch to answer additional questions and clarify ambiguities.

Online workers

We recruited online coders from mTurk, an online labor market developed by Amazon.com to facilitate work that requires human intelligence. The mTurk system allows employers to create small tasks for workers to complete. Workers can select from among thousands of widely-varying jobs.[7]

It is impossible to train online workers interactively because communication occurs only through the mTurk website or by email. In theory one could pay workers to complete extensive online training, but designing systems to enforce attention to the training would be difficult or

[5]Appendix A12 discusses other approaches to measuring latent content.

[6]Data collection was approved by the University of Virginia IRB, project number 2015031700.

[7]mTurk is the most prominent platform; other vendors offer similar services, including some that aggregate multiple worker pools (Vakharia and Lease 2015).

impossible. Instead, we evaluate an alternative: shorter, standardized instruction of many lightly-screened workers. We required workers to be US residents who had completed at least 100 tasks with a 95 percent approval rate. We also required them to complete a background survey, read the coding guidelines, and verify they could watch the videos. Overall, 526 mTurk workers coded at least one ad for us. The average worker coded 53 ads, though the distribution is highly skewed: many dropped out after a few, some coded dozens, and a few coded hundreds.[8]

Comparing online and traditional coders

Both workforces received the same written coding manual; beyond this their training differed substantially. We knew the research assistants and provided comprehensive face-to-face training and discussion to tune their intuitions and refine and align their understanding of each coding category. The online workers were unknown to us and less carefully selected. Our directions aimed to cue the right concepts and alert them to important distinctions, but we relied heavily on their intuitive understanding of concepts.

We analyze the 20 coding decisions for each ad that we describe above (see also Table 1). All but 65 ads were coded by five or more mTurk workers; of those 198 were coded by at least 20 mTurk workers to allow more detailed reliability analysis. 1,512 ads were also coded by research assistants, with 300 double-coded by two research assistants and 85 by all six. Over the course of the project we adjusted coding instructions somewhat, so online coders encountered some variation in instructions and coding items across several distinct waves of coding.[9]

Reliability

We measure inter-coder reliability with Krippendorff's α (Krippendorff 1970; Gwet 2014), which measures the degree to which different coders make the same categorizations, adjusted for chance agreements. It is an extension of Cohen's (1960) canonical kappa statistic that handles ordinal and interval data and multiple non-unique raters.[10]

Table 2 presents reliability for the different types of coding, separately among research assistants and mTurk workers, plus absolute and percentage differences between the two types of workers.[11] The first column displays inter-rater agreement among the research assistants; there is substantial variation across items. Reliability is highest for whether candidates appear in the ad (average alpha is 0.90), followed by the presence of an American flag (0.71) and the optimistic or pessimistic tone of any economic appeals (0.68).[12] Reliability is lower for the presence of economic appeals (0.54) and trait attributions (0.40 for traits attributed to the favored candidate; 0.41 for opponents). Emotional appeals coding is the least reliable, at 0.31. In case the three-way coding of emotion (strong, weak, or no appeal) lowered reliability, we collapsed each to the presence (strong or weak) versus absence of the emotion. This had no effect: reliability on these items also averaged 0.31.

Although some of these coefficients are low, they are consistent with WMP's own coding. Though we could not obtain reliability statistics for their 2010 coding, WMP provided kappa statistics for

[8]See Appendix A3 for more information on the composition of our workforce and our interactions with them and A7 for analyses showing that practice had little impact on coding quality.

[9]We collapse coding waves in our analyses, having found no evidence of systematic variations in reliability or validity; see Appendix A9.

[10]Note that our interest is in *relative* reliability of different coder populations, not the absolute reliability *level*. Appendix A4 shows that our results are unchanged when we use alternate reliability measures.

[11]Analyses conducted in Stata with user-contributed `kappaetc` (Klein 2018). Appendix A4 presents item-level reliability statistics.

[12]One might expect near-perfect reliability on the presence of the flag, which appears simple and manifest. As we discuss below, this item illustrates the striking subtlety of many political advertisements.

Table 1. Coding items

Item	Description	Categories
Manifest Items		
1 Flag	Does an American Flag appear?	yes | no
2 FC Appearance[a]	To what extent is candidate mentioned or shown?	Picture, video, or audio | Actual name | In 'paid for' only | No reference
3 OC Appearance		Picture, video, or audio | Actual name | Vague or generic only | No reference
Pattern Items		
4 Economic appeal	Does ad include an economic appeal?	yes | no
5 Optimistic economic	Is economic appeal optimistic?	yes | no
6 Pessimistic economic	Is economic appeal pessimistic?	yes | no
Latent Items		
7 Enthusiasm 8 Fear 9 Anger 10 Disgust	Does the ad make an appeal to any of the following emotions?	Strong appeal | Weak appeal | No appeal
11 FC Competence	Competence	no | competent
12 OC Competence		no | incompetent
13 FC Leadership	Leadership	no | strong leader
14 OC Leadership		no | weak leader
15 FC Integrity	Integrity	no | has integrity
16 OC Integrity		no | lacks integrity
17 FC Empathy	Empathy	no | empathetic
18 OC Empathy		no | unempathetic
19 FC Ideology	Where would you place each candidate based on this ad?	slider: "very liberal" → "very conserv."
20 OC Ideology		

FC, favored candidate; OC, opposing candidate.
[a]Appearance categories varied slightly; analysis based on dichotomous version for any appearance versus none.

several comparable items in 2014, including manifest (flag) and projective (fear, enthusiasm, anger) items. Across these our research assistant reliability is comparable to theirs.[13]

The power of aggregation: meta-coder reliability gains

The second column of Table 2 shows the corresponding inter-rater reliability statistics among mTurk workers. In short, they are less reliable. Yet the efficiencies of the online labor market enabled us to have each ad coded repeatedly by different coders. If the lower reliability is due to greater random error in the coding by online workers, then we can reduce that error by averaging together multiple coders. As Benoit et al. (2016, 279) point out, "as long as crowd workers are not systematically biased in relation to the 'true' value of the latent quantity of interest … the central tendency of even erratic workers will converge on this true value as the number of workers increases." In this section we exploit the mathematical properties of aggregation to ask whether employing many of the plentiful crowd workers can produce reliability on par with the scarcer student coders. The answer is a clear "yes."

For this analysis we rely on the ads coded by 20 online workers. From those 20 workers, we created a set of four aggregates, each constructed by averaging five randomly-chosen workers. We call each aggregate a "meta-coder," because although it is based on multiple actual coders, we treat it as a single coding for reliability analysis.[14] We rounded each average to produce the same set of outcome values (2, 3, or 101, depending on the item), in order to make the analysis comparable.[15] (As we discuss below, we do not advocate this rounding in actual practice,

[13]Email communication, 2/17/2017.
[14]In Appendix A5, we show that 4-5 coders reduces measurement error substantially.
[15]In Appendix A4 we use mean-squared variation across coders to avoid rounding; the results are equivalent.

Table 2. Inter-coder reliability

	Research assistants	mTurk workers	mTurk versus RA	mTurk versus RA (%)
Flag appears	0.71	0.52	−0.19	−26
Average for candidate appears	0.90	0.83	−0.06	−7
Economic appeal	0.54	0.37	−0.17	−32
Average for economic tone	0.68	0.58	−0.10	−15
Average for emotions	0.31	0.36	+0.05	+16
Average for dichotomized emotions[a]	0.31	0.37	+0.06	+18
Average for FC traits	0.40	0.34	−0.06	−16
Average for OC traits	0.41	0.31	−0.10	−25
Average for ideology	0.63	0.41	−0.23	−36

Krippendorff's α for multiple raters; with ordinal weights for three-point emotions and quadratic weights for 101-point ideology, averaged across individual items.
[a]Emotion coding is on a three-point scale (strong, weak, none); dichotomized versions collapse strong and weak.

preferring to harness the information about the strength or clarity of the underlying stimulus provided by the unrounded average.)

Table 3 presents the aggregation gains by examining agreement among these four meta-coders. The first column summarizes the individual mTurk coder-level reliability; these differ slightly from Table 2's second column because they are calculated across the subset of ads coded by 20 workers. Table 3's second column reports inter-rater agreement among the four mTurk meta-coders; the third column reports the aggregation gain in agreement this produces. The results are quite clear and consistent: aggregation increases inter-coder reliability substantially, increasing alpha by +0.06 to +0.24.[16]

The fourth column shows agreement among the research assistants (for this subset of ads). The final two columns compare mTurk meta-coders with research assistants, in absolute and relative terms. Here, too, the results are relatively consistent and positive: in all areas except the presence of economic appeals and ideology, the mTurk meta-coders achieve higher—and often substantially higher—rates of agreement than research assistants. Meta-coder alpha is 0.05–0.26 higher, an improvement of 5–97 percent. For ideology, the meta-coders are equivalent to research assistants; only on the presence of economic appeals do the meta-coders fall short of the research assistants' reliability, with alpha that is 0.14 lower.

Item-level variation

In Figure 2 we show reliability information for every item, comparing research assistants, individual mTurk workers, and aggregated mTurk meta-coders. Research assistant reliability—corresponding to the fourth column in Table 3—is plotted on the x-axis. mTurk worker reliability is on the y-axis. The arrows depict the gain in reliability when moving from individual mTurk workers (at the base of the arrows, corresponding to the first column of Table 3) to aggregated meta-coders (the arrowheads, corresponding to the second column of Table 3). This figure shows that coding reliability and the benefits of aggregation are not evenly distributed across the items.

First, manifest items: research assistants and individual mTurk workers have high reliability for candidate appearance and aggregation provides modest benefits. On the presence of a flag research assistants are quite reliable, though perhaps lower than we would expect.[17] We examined a number of ads to see what was going on, as we expected this to be a relatively straightforward coding decision. Several things generated disagreement among coders: first, some ads showed a small, partial flag *very* briefly—less than half a second. Other ads included items of clothing (e.g., a construction hard-hat or a necktie) that plausibly—but not certainly—depicted a flag. Finally, one ad included a prominent black-and-white image with a flag in the background. In this case,

[16]Benoit et al. find that Bayesian scaling models produce very similar results to averaging.
[17]The flag item was not coded by 20 mTurk workers, so we cannot aggregate into multiple meta-coders.

Table 3. Reliability gains from aggregation

	mTurk workers (on meta subset)	mTurk meta-coders	Difference: meta-coder gain	Research assistants (on meta subset)	meta-mTurk versus RA	meta-mTurk versus RA (%)
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
Average for economic tone	0.60	0.71	+0.10	0.66	+0.05	+7
Average for emotions	0.29	0.47	+0.18	0.28	+0.19	+67
Average for dichotomized emotions[a]	0.31	0.50	+0.19	0.27	+0.23	+87
Average for FC traits	0.32	0.56	+0.24	0.41	+0.15	+38
Average for OC traits	0.33	0.49	+0.15	0.38	+0.11	+29
Average for ideology	0.43	0.68	+0.24	0.66	+0.01	+2

Krippendorff's α for multiple raters; with ordinal weights for three-point emotions and quadratic weights for 101-point ideology, averaged across individual items.
Meta-coders are average of five randomly-selected coders, rounded to generate a categorical code. Analysis restricted to ads with multiple meta-coders.
[a]Emotion coding is on a three-point scale (strong, weak, none); dichotomized versions collapse strong and weak.

our coding guide may have contributed to the confusion by specifically mentioning "white stars on blue field with red-and-white stripes." Apparently some coders took this to mean that black-and-white images do not count, while others took our instructions as we intended, to signal that only actual *American* flags should be coded.[18] For these ads coders were often about evenly divided. As we discuss in the conclusion of the paper, an aggregated code of 0.5 might actually be a reasonable and valid measure of the flag's prominence in these ads. In any case, these cases illustrate the density of imagery included in ads and the amazing variation in how political messages and symbols are conveyed.

For the two pattern items on economic tone, aggregation increases alpha slightly, such that aggregated mTurk workers and research assistants are similarly and highly reliable. Oddly, aggregated mTurk workers underperformed relative to research assistants at detecting economic appeals in the first place. For this item we asked coders to distinguish discussion of the budgetary or economic impact of government spending (which counts as economic) from more generic references to government spending, or to the impact of spending on citizens rather than the economy (which did not count). We suspect that some online coders did not fully internalize this rather subtle distinction.

Turning to latent items, there is considerable heterogeneity across items; however, on the whole the aggregation gains are large. Ideology demonstrates the highest reliability among both mTurk workers and research assistants, and thanks to very large aggregation gains, the mTurk meta-coders and research assistants achieve similar reliability. This suggests that ideology—despite its complexity as a cognitive and political concept—is frequently communicated in clear-enough terms for coders to achieve reasonably high agreement.

Across the emotions, individual mTurk workers achieve similar reliability to research assistants on average—somewhat better for fear and disgust, and somewhat worse for enthusiasm and anger. There are large aggregation gains here, with alpha increasing by +0.18 across all four emotions (+0.19 when dichotomized). These gains are especially pronounced for disgust, fear, and enthusiasm, leading aggregated meta-coders to outperform research assistants on all but anger.

For traits, aggregation produces gains are also large: they average +0.24 for favored candidate traits and +0.15 for opponents. The gains are most pronounced for the favored candidate's competence and leadership, but are relatively substantial across the board. Although aggregated meta-coders essentially equal or outperform research assistants across the traits and emotions, the absolute reliability levels are moderate at best, mostly falling in the range from 0.4 to 0.6. By definition,

[18]Appendix A8 contains images of these examples.

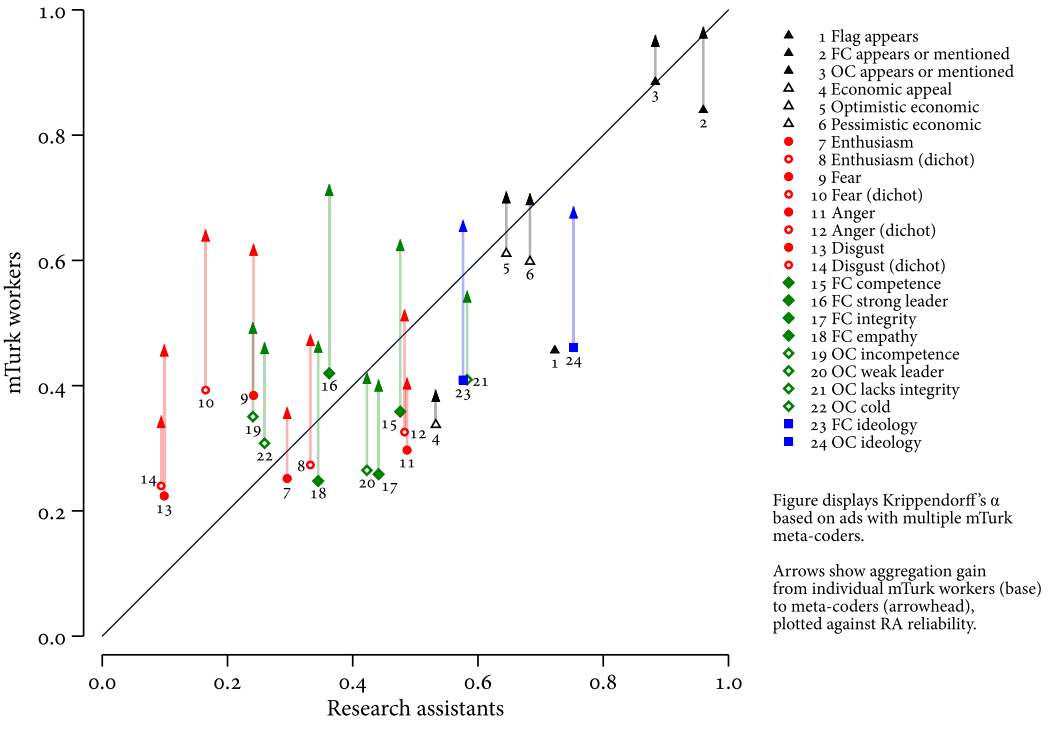

Figure 2. Reliability comparisons by coding item

this means that coders—even after aggregation—often disagree on the presence of traits and emotions. Perhaps individual coders are picking up different aspects of these concepts, in which case the aggregated meta-coder would serve as a useful measure of a complex reality. Alternatively, these concepts may be too subtle and varied for reliable coding. Ultimately this is a question of validity.

Validity

Assessing validity is more difficult than reliability. Ideally, we would compare our coding with the objective truth about each ad. But of course we do not have that truth, and a full-scale validity analysis of each coding category is beyond the scope of this paper. Happily, our interest is not in establishing *absolute* validity; rather, we need to compare the *relative* validity achieved by our two workforces. Our results parallel those for reliability: coding by mTurk workers is less valid than by research assistants, but meta-coders achieve validity that equals and often greatly exceeds that of the research assistants.

We compare the criterion validity of the coding by the two workforces (Carmines and Zeller 1979). We identify external measures of content in the ads (or of the candidates featured in or sponsoring them) that should correlate with our coding. Most of these criterion measures come from the data that WMP provides, which is not completely valid itself. Nevertheless, they represent the best measurement we have of most ad content. Moreover, these data are widely used to study political communication and its effects; as such, they represent an important validity standard.[19] One exception is ideology, where have a truly external validity measure: the DW-NOMINATE estimate of representatives' ideological ideal point based on roll-call voting, available for candidates who served in Congress at some point (Poole and Rosenthal 2007; Bonica 2016).

[19]See Appendix A11 on the use of WMP data to measure political communication.

Our measure of validity is the correlation between each criterion variable and the corresponding coding. We compare these validity coefficients for research assistants, individual mTurk workers, and meta-coders. For example, for economic appeals, the criterion variable is based on WMP measures that suggest economic content, including whether the ad touches on taxes; the deficit, budget, or national debt; government spending; recession or economic stimulus; employment or jobs; and general economic references; and also whether the ad mentions "Main Street" or "Wall Street."

This criterion variable correlates 0.58 with research assistant coding of economic mentions—a solid indication of accurate coding, especially considering that the WMP variables do not capture all possible economic matters. mTurk workers produce lower validity here, with a coefficient of 0.42 ($p < 0.05$ for the difference between the two correlations). However, the meta-coders do better, with validity of 0.56—indistinguishable both substantively and statistically from research assistant validity.

We devised 15 criterion variables that cover 12 of our 20 coding categories. They are detailed in Table 4, which presents the results for each of the 15 validity comparisons. The third column presents the validity coefficients (i.e., correlations between criterion and coding) for research assistants. These range from a high of +0.88, for the appearance of the favored candidate in the ad (the criterion here is simply the WMP coding of the same thing), to +0.22, for coding of claims that the opponent lacks integrity (the criterion is WMP indicating that the ad mentions "corrupt" or "dishonest"). The fourth column shows the difference between the research assistant and mTurk worker validity. In all cases but one, the mTurk validity is lower. Eleven of these 14 decreases are statistically significant, and ten of them exceed 0.05.

The next column shows how aggregation improves validity. As we saw above, coding of economic appeals by mTurk meta-coders is just as valid as the research assistants' coding. In ten of 15 cases, the mTurk meta-coders produce higher validity than research assistants; eight of these differences are statistically significant at $p < 0.05$, and three are quite large (at least +0.10). In five cases mTurk meta-coding is less valid, substantially so on attributions of integrity or lack thereof.

We draw particular attention to ideology, as these involve our best criterion measure. The research assistants are quite valid, with a correlation between ideology coding and DW-NOMINATE scores of 0.58 for favored candidates and 0.62 for opponents. mTurk workers do worse (by −0.15 and −0.16 for favored and opposing candidates, respectively). Once aggregated, however, the correlation is indistinguishable from that among the research assistants. In sum, across all 15 tests the meta-coders and research assistants are essentially equally valid.

Validity variation

Because we are limited by the available criterion measures, we cannot paint a comprehensive validity picture. Nevertheless, we make some observations. Compared with the research assistants, aggregated mTurk meta-coders came slightly closer to the criterion measures on the manifest flag and candidate appearance items. The two types of coders achieved equal validity for the one pattern-based validity test: the presence of economic appeals. Across the latent content, there is systematic variation: the two groups are equally valid for ideology coding; mTurk meta-coders do better across all of the emotional validity tests; and meta-coders do worse on two of three trait tests.

Resources

It cost us about $0.60 to have each ad coded by five mTurk workers.[20] Research assistant wages averaged $0.56 per ad coding. Thus, the two workforces have comparable costs, assuming we have each ad coded by four or five online workers. The online workforce is substantially faster. At this compensation rate, workers coded one ad per minute over four days. At this pace, we could code 2,000 ads five times each in about a week—far faster than any reasonable team of student coders.

[20]See Appendix A10 for more detail on compensation.

Table 4. Validity

Item	Criterion variable	Correlation in RA coding	Relative validity of		Number of ads
			mTurk individual	mTurk-meta	
Flag appears	WMP: American flag appears	0.66	−0.09*	+0.06*	474–1,492
FC appears or mentioned	WMP: FC mentioned or pictured	0.70	−0.01	+0.05*	1,485–1,510
OC appears or mentioned	WMP: OC mentioned or pictured	0.88	−0.02*	+0.03*	1,485–1,510
Economic appeal	WMP: Economic issue or mention	0.58	−0.18*	−0.01	1,416–1,511
Emotion: enthusiasm	WMP: Enthusiasm appeal	0.41	−0.04	+0.10*	1,485–1,510
Emotion: enthusiasm	WMP: Uplifting music	0.42	−0.04	+0.10*	1,485–1,510
Emotion: fear	WMP: Fear appeal	0.38	−0.11*	+0.02	1,485–1,510
Emotion: fear	WMP: Ominous or tense music	0.24	+0.03	+0.16*	1,485–1,510
Emotion: anger	WMP: Anger appeal	0.55	−0.14*	+0.06*	1,415–1,506
Emotion: anger	WMP: Ominous or tense music	0.45	−0.09*	+0.08*	1,415–1,506
FC strong leader	WMP: Ad mentions "tough," "fighter," "experienced"	0.32	−0.10*	−0.00	1,202–1,338
FC integrity	WMP: Ad mentions "honest"	0.24	−0.15*	−0.10*	1,204–1,336
OC lacks integrity	WMP: Ad mentions "corrupt," "dishonest"	0.22	−0.13*	−0.09*	897–1,049
FC ideology	FC DW-NOMINATE ideology score[a]	0.58	−0.15*	+0.03	477–507
OC ideology	OC DW-NOMINATE ideology score	0.62	−0.16*	−0.07	215–256
Average		0.48	−0.09	+0.03	

Restricted to ads with RA coding. Relative validity shows the *difference* in the correlation, compared with that observed in the RA coding. ** $p < 0.01$, * $p < 0.05$ for the comparison with RA reliability.
[a]First-dimension DW-NOMINATE score; available for candidates who served in Congress. Ideology analyses restricted to candidate-sponsored ads.

Discussion

Additional advantages of multiple coding

We have stressed that aggregation reduces measurement error. Multiple coding brings other important benefits. Though most of our coding decisions are binary or ternary, the underlying concepts are continuous. For example, though we code for the presence or absence of integrity, an ad can make more or less strong, direct, and explicit claims about a candidate's honesty. Similarly, coding emotional appeals as absent, weak, or strong captures only some of the wide variation in their strength.

If coders vary in their sensitivity to the presence of a concept, reliability suffers even if they completely agree on its nature. Working to align coders' sensitivity thresholds is statistically inefficient, however, as this discards information about the strength or clarity of a concept within the ad. We can harness that information to produce more valid measures when multiple workers analyze each item, because they provide multiple indicators of the strength and clarity—not just presence—of the concept. As long as ads are randomly assigned to coders this variation produces an unbiased and more valid estimate of the concept.

In addition, we can use multiple measurements of each item to model statistically the interpretive task of coding. These data lend themselves to multi-level modeling, with coding decisions at the lowest level, nested within several higher levels: the ad, the candidate, and the particular race. This could even be extended to a focus on the individual coders, in a non-nested model with coding decisions simultaneously grouped by ad and by coder.

Transparency and replicability

Any scientifically respectable content analysis provides a codebook that documents the coding system. Nevertheless, reproducibility remains a major challenge. Developing guidelines, training

coders, and implementing an analysis involve interaction and discussion. We can never truly replicate this process, because no written codebook can fully document the discussions and other interactions involved. Here the limits of online training are virtues, because this training is entirely standardized. Once the codebook is finalized, it completely documents the coding *and training* system.

Moreover, an open codebook of this sort facilitates "agile" data collection, in which we customize coding systems for each project (Benoit et al. 2016). In addition, we can even vary coding guidelines and training materials experimentally. This permits the systematic study of different training materials on coding quality and provides empirical evidence for weighing trade-offs and guiding choices about procedures.

The nature of human categorization and the wisdom of the crowd

Content analysis sounds deceptively straightforward. Researchers formulate a concept of interest and communicate it to coders, who search for indications of the concept in a body of material. But defining and communicating precise definitions for most concepts of interest is inherently challenging for reasons having to do with the fundamental nature of concepts and the cognitive process of categorization. We think an online system employing numerous coders and an open codebook offers particular advantages for such challenging coding.

The classical theory of conceptual categories—corresponding to everyday understanding—holds that categories are "structured mental representations that encode a set of necessary and sufficient conditions for their application" (Gelman and Wellman 1999; see also Lakoff 1987, 5–11). Our coding included manifest categories of this sort. For example, the category "American flag" is readily described in terms of necessary and sufficient conditions (13 red and white stripes; 50 white stars on a blue field; and so on). Although the ads presented ambiguous cases, in principle we could develop objective rules to cover these variations.

But many categories in life, in content analysis, and in politics are not defined by necessary and sufficient conditions. Wittgenstein uses the concept of a "game" to illustrate this (1953). Games, he points out, share no set of common features: some involve luck, others skill, others both. Some have winners and losers, others not. Moreover, the advent of video games expanded the concept (Lakoff 1987, 16). Categories of this kind are unified not by rules but by "family resemblances." Different games are linked together like "the various resemblances between members of a family: build, features, colour of eyes, gait, temperament, etc. etc. (that) overlap and criss-cross" (Wittgenstein 1953; quoted it Khatchadourian 1966, 206). The concept "game" exists not as a list of rules, but rather as a *cluster* "of gamey attributes, only some of which are instantiated by any one game" (Armstrong et al. 1999, 229).

Many concepts we seek to identify in political communication share this clustered structure. Consider anger. Anger is expressed many different ways—Lakoff spends 36 pages cataloging the synonyms, abstract metaphors, and prototypical scenarios that can convey anger just in language, leaving aside audiovisual imagery (1987, 380–415). It would be impossible to construct a comprehensive coding manual for the concept of anger, or for other clustered concepts. Instead, we traditionally supplement the written rules with intensive training in hopes that coders will "know it when they see it" (in Justice Stewart's famous words about obscenity). But given the unpredictability of this training process, written coding documentation cannot fully capture the training as implemented, and future researchers cannot replicate it exactly. This contributes, we believe, to the well-documented variation in estimates of the same concepts produced by different content analysis teams. For example, Geer (2006, 37) describes dramatic variation in coding of presidential campaign-ad negativity—a *relatively* straightforward concept—with estimates varying from 24 to 54 percent. Scholars hoping to replicate or extend these analyses can train research assistants using the respective coding guides, but they cannot produce in the minds of their coders an understanding that exactly matches that of the original coders.

The open-codebook strategy does not try to align fully each coder's understanding of the concept. Rather than stamping out variation among coders' intuitions, we instead embrace it, or at least tolerate and measure it. We think this approach is especially useful for coding complex, cluster-based concepts for which most people have an intuitive understanding—Potter and Levine-Donnerstein's (1999, 260) "primitive concepts."

We note the greater success here for emotions than for traits. We suspect this is because our coders had richer and more consistent intuitions about emotions than about traits. Scholars of emotion confirm that ordinary people have a well-developed intuitive understanding of major emotions like fear, anger, enthusiasm, and disgust. There are distinct facial expressions, body language, tone of voice, and words associated with each, and people are quite skilled at decoding them (e.g. Frijda 1988). Each emotion involves visceral reactions that everyone has experienced, that coders can draw upon to help recognize them. In contrast, traits like competence, strong leadership, integrity, and empathy are more abstract and context-dependent. What constitutes competence, for example, depends on the task at hand, and some might (reasonably) believe that integrity, strong leadership, and empathy *constitute* competence for a politician. This makes it difficult to distinguish among them in campaign ads. We tried to delineate distinctions among traits; for example, we directed that integrity must include reference to honesty and that "strong values" or "authenticity" do not count by themselves. However, the lower validity on traits among mTurk workers hints that these distinctions may be too far removed from people's primitive concepts to allow online workers to code them well.

Conclusion

In closing, we offer four general recommendations for coding political video using online labor markets. First, have each video coded multiple times, to maximize reliability and validity through aggregation. This repeated coding also produces continuous measures from easy-to-make binary coding decisions, and opens opportunities for modeling the coding process itself.

Second, develop, test, and revise the coding system in person with a small team of experts or research assistants before beginning online coding. We learned this the hard way. Initially we tried to separate "sociotropic" economic appeals that reference the economy from "pocketbook" appeals that touch on the viewer's personal finances (Kinder and Kiewiet 1981). Reliability was terrible; in training the research assistants, and then reviewing many more ads, we learned why: actual campaign ads often mix these appeals in ways that make them very hard to distinguish. Therefore, we dropped the distinction in later rounds of coding. Our experience here—and with trait coding—suggests that scholars should not assume that theoretical concepts map neatly onto real-world political communication.

Third, cue general schema in the coding guide rather than providing exhaustive definitions for concepts. Long lists of criteria and examples are inevitably incomplete, and workers will not attend fully to them anyway. Here we depart from typical content analysis practice, where in-person interaction proposes to fine-tune coders' intuitions. In the online context, it is better to bring the primitive concept to mind and to explain only very important nuances. From there, embrace the inevitable minor variation among coders and rely on multiple measurements to aggregate and/or model that variation.

Finally, implement a custom qualification for online coders that collects demographic and other background information in addition to providing the coding guide. Making workers take this initial step likely excludes coders uninterested in making a good faith coding effort, and incentivizes continued coding work (as workers are invested in completing the initial survey). It also allows the researcher to revoke the qualification of those few workers who systematically ignore instructions.

In summary, this paper provides evidence that workers recruited from online labor markets can classify political content—ranging from manifest to latent—with reliability, validity, and

cost similar to that of traditional research assistants. What is more, these online workers save time, while also making the entire coding process more transparent, replicable, and easier to extend and adapt in future work. And finally, crowd-sourced coding addresses the challenge of coding audio-visual content, helping political analysts identify received meaning through the haze of ambiguous messaging.

Supplementary Material. The supplementary material for this article can be found at https://doi.org/10.1017/psrm.2019.4

Author ORCIDs. Nicholas J. G. Winter, 0000-0002-9121-4746.

Acknowledgements. For their feedback and advice we thank Andrew Clarke, Lindsey Cormack, Paul Freedman, Erika Fowler, Dan Gingerich, Thomas Gray, Tom Guterbock, Jon Kropko, Elizabeth Schwenzfeier, Abby Stewart, David Winter, Sara Winter, and Baobao Zhang. We benefitted from feedback at presentations to the UVa Data Gathering Methodology Study Group and the 2016 Annual Meeting of the American Political Science Association. This project was supported by the Office of the Vice President for Research, the College and Graduate School of Arts & Sciences, and the Quantitative Collaborative at the University of Virginia. Some of the data were obtained from the Wesleyan Media Project, a collaboration between Wesleyan University, Bowdoin College, and Washington State University, and includes media tracking data from Kantar/Campaign Media Analysis Group. The Wesleyan Media Project was sponsored in 2010 by grants from the Sunlight Foundation and The John S. and James L. Knight Foundation. The opinions expressed in this article are those of the authors and do not necessarily reflect the views of the Wesleyan Media Project, the Sunlight Foundation, Knight Foundation or any of its affiliates. This research was conducted before Hughes joined the Pew Research Center and is not based on Pew Research Center data.

References

Armstrong SL, Gleitman LR and Gleitman H (1999) What some concepts might not be. In Margolis E and Laurence S (eds). *Concepts: Core Readings*. Cambridge, MA: MIT Press, pp. 225–59.

Benoit WL (2010) Content analysis in political communication. In Bucy EP and Holbert RL (eds). *Sourcebook for Political Communication Research: Methods, Measures, and Analytic Techniques*. New York: Taylor & Francis, pp. 268–79.

Benoit K, Laver M and Mikhaylov S (2009) Treating words as data with error: uncertainty in text statements of policy positions. *American Journal of Political Science* **53**, 495–513.

Benoit K, Conway D, Lauderdale BE, Laver M and Mikhaylov S (2016) Crowd-sourced text analysis: reproducible and agile production of political data. *American Political Science Review* **110**, 278–95.

Bonica A (2016) *Database on Ideology, Money in Politics, and Elections: Public Version 2.0*. [Computer file]. Stanford, CA: Stanford University Libraries. Available at http://data.stanford.edu/dime.

Brader T (ed) (2006) *Campaigning for Hearts and Minds: How Emotional Appeals in Political Ads Work*. Chicago: University of Chicago Press.

Budak C, Goel S and Rao JM (2016) Fair and balanced? Quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly* **80**(S1), 250–71.

Carmines EG and Zeller RA (1979) *Reliability and Validity Assessment*. Thousand Oaks, CA: Sage.

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

Fowler EF, Franz MM and Ridout TN (2014) *Political Advertising in 2010*. [Computer file]. Version 1.3. Middletown, CT: The Wesleyan Media Project; Department of Government at Wesleyan University.

Frijda NH (1988) The laws of emotion. *American Psychologist* **43**, 349–58.

Geer JG (2006) *In Defense of Negativity: Attack Ads in Presidential Campaigns*. Chicago: University of Chicago Press.

Gelman SA and Wellman HM (1999) Insides and essences: early understandings of the non-obvious. In Eric M and Laurence S (eds). *Concepts: Core Readings*. Cambridge, MA: MIT Press, pp. 613–37.

Grabe ME and Bucy EP (2009) *Image Bite Politics: News and the Visual Framing of Elections*. New York: Oxford University Press.

Graber DA (1987) Television news without pictures? *Critical Studies in Mass Communication* **4**, 74–8.

Graber DA and Smith JM (2005) Political communication faces the 21st century. *Journal of Communication* **55**, 479–507.

Grimmer J and Stewart BM (2013) Text as data: the promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis* **21**, 267–97.

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

Hayes D (2011) When gender and party collide: stereotyping in candidate trait attribution. *Politics & Gender* **7**, 133–65.

Khatchadourian H (1966) Common names and "Family Resemblances". In *Wittgenstein: The Philosophical Investigations*, George P (ed). Garden City, NY: Anchor Books, pp. 203–30.

Kinder DR and Kiewiet. DR (1981) Sociotropic politics: the American case. *British Journal of Political Science* **11**, 129–61.

Kinder DR, Peters MD, Abelson RP and Fiske ST (1980) Presidential prototypes. *Political Behavior* **2**, 315–37.

Klein D (2018) Implementing a general framework for assessing interrater agreement in stata. *The Stata Journal* **18**, 871–901.
Krippendorff K (1970) Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement* **30**, 61–70.
Lakoff G (1987) *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind*. Chicago: University of Chicago Press.
Lind F, Gruber M and Boomgaarden HG (2017) Content analysis by the crowd: assessing the usability of crowdsourcing for coding latent constructs. *Communication Methods and Measures* **11**, 191–209.
Mann R (2011) *Daisy Petals and Mushroom Clouds: LBJ, Barry Goldwater, and the Ad That Changed American Politics*. Baton Rouge: Louisiana State University Press.
Masters RD and Sullivan DG (1993) Nonverbal behavior and leadership: emotion and cognition in political information processing. In Shanto I and McGuire WJ (eds). *Explorations in Political Psychology*. Durham: Duke University Press, pp. 150–82.
Mendelberg T (2001) *The Race Card: Campaign Strategy, Implicit Messages, and the Norm of Equality*. Princeton, NJ: Princeton University Press.
Messaris P (1997) *Visual Persuasion: The Role of Images in Advertising*. Thousand Oaks, CA: Sage Publications.
Messaris P and Abraham L (2001) The role of images in framing news stories. In Stephen DR, Gandy OH and Grant AE (eds). *Framing Public Life: Perspectives on Media and Our Understanding of the Social World*. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 215–26.
Poole KT and Rosenthal H (2007) *Ideology & Congress*. 2nd rev. edn. New Brunswick: Transaction Publishers.
Potter WJ and Levine-Donnerstein D (1999) Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research* **27**, 258–84.
Reynolds TJ and Whitlark DB (1995) Applying laddering data to communications strategy and advertising practice. *Journal of Advertising Research* **35**, 9–17.
Rosenberg SW, Bohan L, McCafferty P and Harris. K (1986) The image and the vote: the effect of candidate presentation on voter preference. *American Journal of Political Science* **30**, 108–27.
Schill D (2012) The visual image and the political image: a review of visual communication research in the field of political communication. *Review of Communication* **12**, 118–42.
Spielvogel C (2005) You know where I stand: moral framing of the war on terrorism and the Iraq war in the 2004 presidential campaign. *Rhetoric & Public Affairs* **8**, 549–69.
Strach P, Zuber K, Fowler EF, Ridout TN and Searles K (2015) In a different voice? Explaining the use of men and women as voice-over announcers in political advertising. *Political Communication* **32**, 183–205.
Vakharia D and Lease M (2015) Beyond AMT: an analysis of paid crowd work platforms. *iConference 2015 Proceedings*. Available at http://hdl.handle.net/2142/73639.
Weber R, Mangus JM, Huskey R, Hopp FR, Amir O, *et al.* (2018) Extracting latent moral information from text narratives: relevance, challenges, and solutions. *Communication Methods and Measures* **12**, 119–39.
Wittgenstein L (1953) *Philosophical Investigations*. Oxford: B. Blackwell.
Woollacott J (1982) Messages and meanings. In Tony B, Curran J, Gurevitch M and Wollacott J (eds). *Culture, Society and the Media*. New York: Routledge, pp. 87–109.

Cite this article: Winter NJG, Hughes AG, Sanders LM (2019). Online coders, open codebooks: New opportunities for content analysis of political communication. *Political Science Research and Methods* 1–16. https://doi.org/10.1017/psrm.2019.4