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

Forthcoming, Political Science Research & Methods

Nicholas J. G. Winter
University of Virginia
nwinter@virginia.edu

Adam G. Hughes
Pew Research Center
aghughes@gmail.com

Lynn M. Sanders
University of Virginia
lms5x@virginia.edu

December 14, 2018

Word count: 8,910
(without title, abstract, & acknowledgements)
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, or complicate the content of textual information. To address these measurement challenges, we describe a fully reproducible approach to analyzing video content using minimally – but systematically – trained online workers. By aggregating the work of multiple coders, the online approach achieves 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.

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. In addition, the paper benefitted from feedback at presentations to the UVa Data Gathering Methodology Study Group and at the 2016 Annual Meeting of the American Political Science Association. Support for this project was provided by the Office of the Vice President for Research, the College and Graduate School of Arts & Sciences, the Political Cognition Laboratory, and the Quantitative Collaborative at the University of Virginia. Some of the data analyzed in this paper 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 in Washington, D.C. 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 author(s) 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.
We are in the midst of a content-analysis revolution driven by computing and crowd-sourcing. New computational and crowd-sourced techniques have helped automate the classification of political texts (Grimmer and Stewart 2013; Benoit et al. 2009; Benoit et al. 2016). 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 overlooked because analyzing it is difficult and time-consuming and because there has been less theorizing on how audiovisual media convey meaning.

In this article 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 perform well at measuring complex audiovisual content. We show that crowd-sourcing is particularly suited to this task, where coding categories are not reducible to objective rules. In addition, because our approach is more reproducible than traditional content analysis, it promises to help scholars refine theories of how audiovisual materials convey meaning.

We recruited a large number of 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 and colleagues (2016) show 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 important, 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 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. Mann argues that 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); in addition, 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 and colleagues 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” (1986, 123).

In addition, pictures and sounds evoke strong emotional responses that shape the processing of accompanying verbal material. Brader shows, for example, that anxiety cues lead viewers to attend to and remember information conveyed by campaign ads (2006); in 2008 the Clinton campaign used exactly these sorts of cues in the “3 a.m.” ad to claim that Obama was not ready to lead in a dangerous world. Irony and humor can also reinforce an ad’s message. For example, in “Tank Ride,” the 1988 Bush campaign paired the claim that Michael 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 footage of John Kerry windsurfing to forge a link between Kerry’s policymaking and his personal character (Spielvogel 2005) In other cases, visuals convey messages that do not appear in the verbal channel at all. This is especially important for racial messages; for example, Mendelberg shows how a mugshot of William Horton in the 1988 “Weekend Passes” ad primed white Americans’ racial resentment without explicit racial references (2001). Thus, concludes Benoit, “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)” (2010, 276).

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, the symbols used to communicate physically resemble the things they represent, in contrast with the arbitrary signs of spoken or written language (Messaris 1997). Iconography produces powerful psychological effects: viewers’ 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).

There is broad agreement among students of political communication that visuals are both important and understudied. For decades scholars have called for more attention to the visual face of political communication. In 1987, Doris Graber concluded that coding only verbal content distorts the meaning of political messages. Almost two decades later, though, Graber and Smith found little evidence that scholars employed visual coding (2005, 492). And in 2012, Schill argued that “the visual aspects of political communication remain one of the least studied and the least understood areas, and research focusing on visual symbols in political communication is severely lacking” (119).

The logistical difficulty of coding video is compounded by the fact that visual communication lacks the propositional syntax that governs formal language; it operates by informal and less well-understood rules. While there are “relatively precise conventions for indicating spatial or temporal relationships among two or more images,” Messaris argues, “visual communication is characterized by a lack of means for identifying other ways in which images might be related to each other” (1997, x). Thus, visual communication cannot convey the precise propositional claims of certain kinds of political argumentation (e.g., this policy is a bad idea because it will lead to outcomes X and Y). Instead, as we discuss above, visual messages are ambiguous, carrying multiple messages simultaneously. Even when conveying spoken language, the speaker’s tone of voice and gender, among other things, can shape powerfully how the message is received (e.g. Strach et al. 2015).

To advance understanding of visual political communication, we need more and better measurement of the content of real-life communications. As long as coding visual content is a time-consuming and difficult-to-replicate enterprise, we cannot pursue the cumulative empirical research needed to support theory
development. By making coding more efficient, open, and replicable, the power of the crowd promises to facilitate both empirical and theoretical development in the study of political communication.

THE ONLINE CODING INTERFACE, THE CODERS, AND OUR CODING TASKS

Video 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 (see figure 1). The interface showed coders a single ad, a data entry form, and an abbreviated summary of the coding instructions. Coders could expand the instructions on-screen and also download the codebook. They viewed each randomly-selected ad one or more times while making their coding decisions.

Figure 1: Coding Interface

---

1 See appendix A1 for information on the implementation of the coding portal.

2 Online workers averaged 72 seconds coding each 30-second ad. Time varied greatly (standard deviation of 52 seconds). See appendix A6 for analysis of variation in coding time and its (lack of) impact on reliability and validity.
The coding tasks

We included a diverse array of coding items. We sought concepts that (1) range in the likely difficulty of coding accurately; (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 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 message. 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” (1999, 259).

Table 1: Coding items

<table>
<thead>
<tr>
<th>Coding item</th>
<th>Label in interface</th>
<th>Coding categories</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flag</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 Flag appears</td>
<td>Does and American Flag appear?</td>
<td>yes/no</td>
<td>manifest</td>
</tr>
<tr>
<td>Candidate appearance in ad1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 FC3 Appearance</td>
<td>To what extent is candidate mentioned or shown?</td>
<td>Picture, video, or audio/ Actual name/ In ‘paid for’ only/ No reference</td>
<td>manifest</td>
</tr>
<tr>
<td>3 OC Appearance</td>
<td>“ ” “ ”</td>
<td>Picture, video, or audio/ Actual name/ Vague or generic only/ No reference</td>
<td>manifest</td>
</tr>
<tr>
<td>Economic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 Economic appeal</td>
<td>Does ad include an economic appeal?</td>
<td>yes/no</td>
<td>pattern</td>
</tr>
<tr>
<td>5 Optimistic economic</td>
<td>Is economic appeal optimistic?</td>
<td>yes/no</td>
<td>pattern</td>
</tr>
<tr>
<td>6 Pessimistic economic</td>
<td>Is economic appeal pessimistic?</td>
<td>yes/no</td>
<td>pattern</td>
</tr>
<tr>
<td>Emotions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 Enthusiasm</td>
<td>Does the ad make an appeal to any of the following emotions/ Enthusiasm</td>
<td>Strong appeal/ Weak appeal/ No appeal</td>
<td>latent</td>
</tr>
<tr>
<td>8 Fear</td>
<td>… Fear</td>
<td>Strong appeal/ Weak appeal/ No appeal</td>
<td>latent</td>
</tr>
<tr>
<td>9 Anger</td>
<td>… Anger</td>
<td>Strong appeal/ Weak appeal/ No appeal</td>
<td>latent</td>
</tr>
<tr>
<td>10 Disgust</td>
<td>… Disgust</td>
<td>Strong appeal/ Weak appeal/ No appeal</td>
<td>latent</td>
</tr>
<tr>
<td>Traits</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11 FC Competence</td>
<td>Competence</td>
<td>no/ competent</td>
<td>latent</td>
</tr>
<tr>
<td>12 OC Competence</td>
<td>“ ”</td>
<td>no/ incompetent</td>
<td>latent</td>
</tr>
<tr>
<td>13 FC Leadership</td>
<td>Leadership</td>
<td>no/ strong leader</td>
<td>latent</td>
</tr>
<tr>
<td>14 OC Leadership</td>
<td>“ ”</td>
<td>no/ weak leader</td>
<td>latent</td>
</tr>
<tr>
<td>15 FC Integrity</td>
<td>Integrity</td>
<td>no/ has integrity</td>
<td>latent</td>
</tr>
<tr>
<td>16 OC Integrity</td>
<td>“ ”</td>
<td>no/ lacks integrity</td>
<td>latent</td>
</tr>
<tr>
<td>17 FC Empathy</td>
<td>Empathy</td>
<td>no/ empathetic</td>
<td>latent</td>
</tr>
<tr>
<td>18 OC Emphy</td>
<td>“ ”</td>
<td>no/ unempathetic</td>
<td>latent</td>
</tr>
<tr>
<td>Ideology</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19 FC Ideology</td>
<td>Where would you place each candidate based on this ad?</td>
<td>[slider with endpoints “very liberal” and “very conservative”]</td>
<td>latent</td>
</tr>
<tr>
<td>20 OC Ideology</td>
<td>“ ” “ ”</td>
<td>[slider with endpoints “very liberal” and “very conservative”]</td>
<td>latent</td>
</tr>
</tbody>
</table>

1Appearance categories varied somewhat across waves; analysis based on collapsed, dichotomous version for “NO appearance” vs. any appearance
2“FC” = favored candidate; “OC” = opposing candidate
3See appendix A2 for additional information about the coding process.
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). The remaining items fall 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). 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.

Latent projective concepts like these present the biggest challenges for content analysis, because they are complex and symbolic and do not have concise, rule-based definitions. Accordingly, reliable and valid coding is generally thought to require extensive face-to-face training. Analysts have deployed several approaches to measure these constructs, including qualitative and interpretive studies (e.g. Oddo 2011; Bell and Entman 2011; Limbert and Bullock 2009); content analysis by carefully-trained undergraduate research assistants (e.g. Bystrom and Dimitrova 2014; Weber et al. 2018); and content analysis by experts (sometimes study authors themselves; e.g. Coleman and Wu 2010), and often as part of validation tests or comparisons to other coding techniques (e.g. Meirick et al. 2018; Gemenis 2015; Carlson and Montgomery 2017). Unsupervised machine learning approaches like topic models are also used to identify latent constructs in text (Nardulli et al. 2015; Young and Soroka 2012), while supervised machine vision models can help classify image content (e.g. Casas and Williams 2017).

However, these approaches have real limits. Qualitative and interpretive approaches are difficult to replicate and generalize. Expert and research assistant coding are time consuming and expensive, and thus cannot keep pace with the massive increases in political audiovisual content. While machine learning has great promise, unsupervised models require human validation and supervised approaches require a large number of human-classified examples from which to learn, which brings us back to the original problem. Moreover, automated image classification struggles especially with subtle, latent constructs. For example, Casas and Williams (2017) used unsupervised machine learning approaches like topic models to identify latent constructs in text (Nardulli et al. 2015; Young and Soroka 2012), while supervised machine vision models can help classify image content (e.g. Casas and Williams 2017).

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).
Williams (2017) trained an automated classifier to identify political protests in images posted to Twitter, along with the emotions evoked by those images. They achieved excellent reliability in identifying images of protests, but the classifier was much less accurate at identifying the latent emotional constructs.

We expect that ordinary people are well-suited to this task. Although these concepts are hard to define precisely, most people understand them intuitively, making them something Potter and Levine-Donnerstein call “primitive concepts” (1999, 260). This is clearly true for emotions. Most American coders share broad understandings of fear, anger, and other emotions, and we expect that coders should be able to identify and distinguish between them, especially if the instructions cue culturally-consistent cognitive schema. Identifying traits might be harder, though here too we believe that ordinary people will have relatively clear understanding. 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. Below we ask whether large numbers of minimally trained coders can draw successfully on this intuitive knowledge to produce reliable and valid coding.

There is some precedent for using the crowd to code for textual latent content. Lind, Gruber, and Boomgaarden (2017) show that the crowd equals traditional coders in identifying the presence and valence of target evaluations in text sentences. Weber and colleagues (2018) demonstrate that ordinary citizens can draw on “moral intuitions” to code text. And Budak and colleagues (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 U.S. 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 assistant coders

We hired undergraduate research assistants with recruitment, training, and working conditions that followed standard practice. We emailed advertisements to political science majors and selected six students we knew to be capable. 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 online codebook, practiced coding several ads together, answered questions, and discussed the coding 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 coders

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 tasks.

Because contact between employers and workers occurs through the mTurk website or by email, interactive training and discussion are impossible. In theory one could pay workers to complete extensive online training, but designing systems to enforce attention to the training would be difficult or impossible. Instead, we instead evaluate an alternative: shorter, standardized instruction of many lightly-screened workers. We required workers to be U.S. 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 see and hear the videos on their computer. 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.

---

5 The data collection was approved by the University of Virginia IRB, project number 2015031700.
6 mTurk was one of the earliest crowd source platforms; other vendors offer similar services, including aggregators that draw on multiple worker pools (Vakharia and Lease 2015).
7 See appendix A3 for more information on the composition of our workforce and our interactions with them. We explored whether validity and reliability change as coders get practice; basically, it does not (see appendix A7).
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 and summarize in 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 RAs 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.

Reliability

We measure inter-coder reliability with Krippendorff’s alpha (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.9

Table 2 presents summary reliability for the different coding items, separately among research assistants and mTurk workers, along with absolute and percentage differences between the two types of workers.10 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

---

8 We collapse coding waves in our analyses, having found no evidence of systematic variations in reliability or validity. See appendix A9 for more information on this variation.
9 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.
10 See appendix A4 for item-level reliability statistics.
appeals (0.68). 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) vs. absence of the emotion. This had no effect: reliability on these items also averaged 0.31.

Table 2: Inter-coder reliability (Krippendorff’s α)

<table>
<thead>
<tr>
<th></th>
<th>Research assistants</th>
<th>mTurk workers</th>
<th>mTurk vs. RA (%)</th>
<th>mTurk vs. RA (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flag appears</td>
<td>0.71</td>
<td>0.52</td>
<td>-0.19</td>
<td>-26%</td>
</tr>
<tr>
<td>Average for candidate appears</td>
<td>0.90</td>
<td>0.83</td>
<td>-0.06</td>
<td>-7%</td>
</tr>
<tr>
<td>Economic appeal</td>
<td>0.54</td>
<td>0.37</td>
<td>-0.17</td>
<td>-32%</td>
</tr>
<tr>
<td>Average for Economic tone</td>
<td>0.68</td>
<td>0.58</td>
<td>-0.10</td>
<td>-15%</td>
</tr>
<tr>
<td>Average for emotions</td>
<td>0.31</td>
<td>0.36</td>
<td>+0.05</td>
<td>+16%</td>
</tr>
<tr>
<td>Average for dichotomized emotions¹</td>
<td>0.31</td>
<td>0.37</td>
<td>+0.06</td>
<td>+18%</td>
</tr>
<tr>
<td>Average for FC traits</td>
<td>0.40</td>
<td>0.34</td>
<td>-0.06</td>
<td>-16%</td>
</tr>
<tr>
<td>Average for OC traits</td>
<td>0.41</td>
<td>0.31</td>
<td>-0.10</td>
<td>-25%</td>
</tr>
<tr>
<td>Average for ideology</td>
<td>0.63</td>
<td>0.41</td>
<td>-0.23</td>
<td>-36%</td>
</tr>
</tbody>
</table>

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.

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 several comparable items in 2014, including manifest (flag) and projective (fear, enthusiasm, anger) items. Across these our research assistant reliability is comparable to theirs.²

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.

---

¹ One might expect near-perfect reliability on the presence of the flag, which seems to be clear-cut manifest content. In practice this item illustrates the striking subtlety of many political advertisements; we discuss this further below.

² Email communication, 2/17/2017.
As Benoit and colleagues 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” (2016, 279). 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.\(^{13}\) 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 fully comparable.\(^{14}\) (As we discuss below, we do not advocate this rounding in actual practice, 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

<table>
<thead>
<tr>
<th>Table 3: Reliability gains from aggregation (Krippendorff’s α)</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="#" alt="Table content" /></td>
</tr>
</tbody>
</table>

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 kappaestc (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.

\(^{13}\) In appendix A5, we show that 4-5 coders reduces measurement error substantially.

\(^{14}\) In appendix A4 we measure reliability as the mean-squared variation across coders to avoid rounding; the results are equivalent.
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.\textsuperscript{15}

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 to 0.26 higher, an improvement of five to 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 reliability and in aggregation gain*

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 arrow heads, 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, and individual mTurk workers rather less so.\textsuperscript{16}

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

\textsuperscript{15} Benoit et al. find that more complex Bayesian scaling models produce very similar results as this sort of simple average.

\textsuperscript{16} The flag item was not included in the round of coding by 20 mTurk workers, so we are not able to aggregate into multiple mTurk meta-coders.
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, 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. 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.

---

17 Screenshots of these examples appear in appendix A8.
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, 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, at least without substantially more training and consultation than we provided even to our research assistants. Ultimately this is a question of validity, to which we turn now.
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. One exception is ideology, where we 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).

We calculate validity coefficients as 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, we construct our criterion variable from several WMP measures that suggest economic content, including whether the ad touches the issues of taxes; the deficit, budget, or national debt; government spending; recession or economic stimulus; employment or jobs; and general economic references; as well as whether the ad mentions “Main Street” or “Wall Street.”

This criterion variable correlates 0.58 with research assistant coding of economic mentions. This is solid reliability, especially because the WMP issue variables do not capture all possible economic matters. mTurk workers produce lower validity here, with a validity coefficient of 0.42 (p<0.05 for the difference

---

18 Scholars use WMP data to study campaign negativity and voter mobilization (Goldstein and Freedman 2002; Krupnikov 2011), issue agendas (Banda and Carsey 2015), issue publics (Sides and Karch 2008; Claibourn and Martin 2012), voter learning (Claibourn 2008; Wolak 2009), campaign persuasion (Hillygus and Shields 2008; Franz et al. 2007), gender in campaigns (Sapiro et al. 2011; Schaffner 2005), and race and ethnicity in campaigns (Abrajano 2010).
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 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 10 of them exceed 0.05.

The next column shows that 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 10 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.

Table 4: Validity Analysis

<table>
<thead>
<tr>
<th>Coding variable</th>
<th>Criterion variable</th>
<th>Correlation in RA coding</th>
<th>Relative validity of mTurk coding</th>
<th>Number of ads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flag appears</td>
<td>WMP: American flag appears</td>
<td>0.66</td>
<td>“0.09” ** “0.06” *</td>
<td>474 - 1,492</td>
</tr>
<tr>
<td>FC appears or mentioned</td>
<td>WMP: FC mentioned or pictured</td>
<td>0.70</td>
<td>“0.01” ** “0.05” *</td>
<td>1,485 - 1,510</td>
</tr>
<tr>
<td>OC appears or mentioned</td>
<td>WMP: OC mentioned or pictured</td>
<td>0.88</td>
<td>“0.02” ** “0.03” *</td>
<td>1,485 - 1,510</td>
</tr>
<tr>
<td>Economic appeal</td>
<td>WMP: Economic issue or mention</td>
<td>0.45</td>
<td>“0.04” ** “0.10” *</td>
<td>1,485 - 1,510</td>
</tr>
<tr>
<td>Emotion: enthusiasm</td>
<td>WMP: Enthusiasm appeal</td>
<td>0.12</td>
<td>“0.02” ** “0.02” *</td>
<td>1,485 - 1,510</td>
</tr>
<tr>
<td>Emotion: fear</td>
<td>WMP: Fear appeal</td>
<td>0.38</td>
<td>“0.12” ** “0.16” *</td>
<td>1,485 - 1,510</td>
</tr>
<tr>
<td>Emotion: fear</td>
<td>WMP: Ominous or tense music</td>
<td>0.24</td>
<td>“0.03” ** “0.16” *</td>
<td>1,485 - 1,510</td>
</tr>
<tr>
<td>Emotion: anger</td>
<td>WMP: Anger appeal</td>
<td>0.55</td>
<td>“0.14” ** “0.06” *</td>
<td>1,415 - 1,506</td>
</tr>
<tr>
<td>Emotion: anger</td>
<td>WMP: Ominous or tense music</td>
<td>0.45</td>
<td>“0.09” ** “0.08” *</td>
<td>1,415 - 1,506</td>
</tr>
<tr>
<td>FC strong leader</td>
<td>WMP: Ad mentions “tough,” “fighter,” “experienced”</td>
<td>0.32</td>
<td>“0.10” ** “0.00” *</td>
<td>1,202 - 1,338</td>
</tr>
<tr>
<td>FC integrity</td>
<td>WMP: Ad mentions “honest”</td>
<td>0.24</td>
<td>“0.15” ** “0.10” *</td>
<td>1,204 - 1,336</td>
</tr>
<tr>
<td>OC lacks integrity</td>
<td>WMP: Ad mentions “corrupt,” “dishonest”</td>
<td>0.22</td>
<td>“0.13” ** “0.09” *</td>
<td>897 - 1,049</td>
</tr>
<tr>
<td>FC ideology</td>
<td>FC DW-NOMINATE ideology score</td>
<td>0.58</td>
<td>“0.15” ** “0.03”</td>
<td>477 - 507</td>
</tr>
<tr>
<td>OC ideology</td>
<td>OC DW-NOMINATE ideology score</td>
<td>0.62</td>
<td>“0.16” ** “0.07”</td>
<td>215 - 236</td>
</tr>
</tbody>
</table>

Analysis among ads for which we have RA coding. Relative validity columns show the difference in the correlation, compared with that observed in the research assistant coding.

** p<0.01, * p<0.05 for the difference between correlation coefficients.

1 First-dimension DW-NOMINATE score; available for candidates who held office at some point. Ideology analyses restricted to candidate-sponsored ads.

17
We draw particular attention to the ideology, as these involve our best criterion measure. The RAs 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 reliability picture. Nevertheless, we make some observations. Compared with the research assistants, aggregated mTurk meta-coders achieved slightly higher reliability 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

Our final comparisons focus on money and time. After some tinkering with compensation rates, it cost us $0.60 to have each ad coded by five mTurk workers, including the Amazon commission. 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.

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

19 See appendix A10 for more detail on compensation.
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 providing 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 modelling, 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. In addition, we can even vary coding guidelines and training materials experimentally. This allow us to study systematically the impact of different training materials on coding quality and to use empirical evidence to weigh trade-offs and guide 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, 10; 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 Potter Stewart’s famous words about obscenity). But given the ad hoc, unpredictable nature 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 describes dramatic variation in coding of presidential campaign-ad negativity—a relatively straightforward concept—with estimates varying from 24 percent to 54 percent (2006, 37). 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 “primitive concepts” (1999, 260).

We acknowledge that we had somewhat 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 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 by 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 on these items 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 makes 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 likely 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 for traditional research assistants. What’s 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.
References


