Analyzing audiovisual communication is challenging because its content is highly symbolic and less rule-governed than verbal material. We describe a fully reproducible approach to analyzing video content that relies on systematically but minimally trained online workers. By aggregating the work of multiple coders, the online approach achieves reliability, validity, and costs that equal that of traditional, intensively-trained RAs, with much greater speed, transparency, and replicability. Elaborating on the unique features of visual communication and the cognitive structuring of complex concepts, we argue that measurement strategies that rely on the “wisdom of the crowd” are particularly well-suited to coding ambiguous and intricate audiovisual political content.
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 contained within contemporary political communications. Visual content is frequently overlooked because analyzing it is difficult and time-consuming and because the ways visual media convey meaning are less well-theorized.

This article provides both practical and theoretical contributions. We describe a transparent, efficient, and reproducible approach to analyzing visual political communications. Focusing on political television advertising—a paradigmatic case of complex visual communication—we demonstrate that crowd-sourced online coders perform well at measuring difficult visual content. This approach promises to catalyze progress in the measurement of visual communication. Our discussion amplifies how crowd-sourcing excels especially for the analysis of complex visual content, where coding categories are often not fully reducible to objective rules.

We recruited a large number of coders from an online labor market and trained them systematically, yet less intensively than typically advised; as one guidebook directs, “three words describe good coder preparation: train, train, train” (Neuendorf 2002, 133). We supplied our coders with an online codebook through a custom-programmed web platform, with which coders trained themselves with no feedback from us. In this paper we compare this approach with a traditional method relying on a small number of research-assistant coders who we trained intensively and interactively. Across a range of coding tasks from concrete and literal to abstract and symbolic, we show that the online workforce achieves similar reliability, validity, and cost as traditional RAs. The individual coding decisions are of lower quality, but also less expensive. Crucially, these lower per-coding costs allow us to classify each piece of content repeatedly. By aggregating multiple coding decisions into “meta-coding,” we achieve reliability and validity comparable to or better than traditional RAs, at comparable net cost.
Using online workers offers several additional advantages. First, they complete coding much faster than any feasible team of traditional RAs, meaning that researchers can flexibly adapt coding protocols to their research question rather than relying on preexisting coding. Second, the repeated coding of each item generates measures of ambiguity and coding uncertainty, which opens up additional avenues for substantive analysis. Third, the standardized, hands-off training ensures that the data creation process is completely open and fully reproducible. This allows scholars to build on each other’s work by facilitating precise replication of coding protocols.

We build on Benoit and colleagues (2016), who showed that aggregated non-experts recruited online are as adept as subject-matter experts at classifying ideology in party manifestos. They also highlighted the advantages of online classification for transparent, reproducible, and agile—i.e., flexible and adaptable—data collection. We extend their work in three ways. First, we compare coding by the crowd not with experts, but with student RAs, who are the typical coders for political communications research. Second, we evaluate coding of extended, real-world political communication. Our coders interpreted complete political texts—advertisements, aired in full—rather than discrete sentences or sound-bites. Third, in addition to ideology, we examine coding of a wide range of concepts that figure heavily in audiovisual communication, and which vary from relatively objective and clear-cut to quite abstract and symbolic. Finally and most important, we focus on the complex medium of video. As we argue below, audiovisual communication is intrinsically difficult to code; therefore it represents an important and demanding test of online coders’ ability to parse meaning from political communication. It also highlights the potential for new technologies to capture real-world perceptions of meaning. Our findings thus speak not just to the analysis of political advertising, but to the full range of audio- and video-based content, such as television and radio news and infotainment, and similarly complex non-political advertising.

The medium and the message

Analysis of visual political communication lags substantially behind work on verbal text. The Enlightenment prioritization of rational argument and the presumption that “verbal arguments are ... the primary conduit of reason” (Grabe and Bucy 2009, 6) is partly to blame. More prosaically,
this neglect is rooted in the conceptual and logistical challenges of theorizing and measuring visual—as opposed to purely text-based—communications. But the visual channel plays a dominant role in human communication and social interaction. Compared with language processing, visual perception is faster, more memorable, and more apt to engage emotion and cognition simultaneously (Grabe and Bucy 2009, 12-21). Visuals both shape the processing of accompanying verbal material and convey information directly, including myriad social cues: people make powerful inferences from gestures and facial displays about traits, emotional states, behavioral intentions and motivations, and more (e.g. Masters and Sullivan 1993; Ekman and Rosenberg 1997; Todorov et al. 2005; Sullivan and Masters 1988).

Two features of visual communications have important implications for analysis (Messaris 1997). First, visuals are *iconographic*—that is, the symbols used to communicate physically resemble the things they represent, in contrast with arbitrary signs of spoken or written language. Iconography produces powerful psychological effects: viewers’ cognitive reactions to visual representations of actions mirrors the reactions of actual participation. Visual images enhance perceived realism, potentially increasing their persuasive power while undermining awareness of persuasive effects. This implies that myriad artistic choices, such as camera positioning and lighting, can powerfully shape how viewers identify with and understand the things they see.¹

Second, visuals lack the propositional syntax that governs formal language. 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, visual messages are ambiguous and polysemous, carrying multiple messages simultaneously. Visuals also excel at evoking emotion and at imparting subtle or implicit meanings, allowing them to provoke associations the speaker might not want to endorse openly, such as implicit racial cues (Mendelberg 2001), or might not be

¹ For example, scholars have shown that camera perspective shapes jurors’ interpretation of videotaped confessions (e.g. Ratcliff et al. 2006; Lassiter et al. 2002).
able to defend fully, such as the association between a president and good economic times. Moreover, the powerful, subtle, and non-rule-based features of visual communication are amplified by the audio channel, which also excels at conveying subtle and emotion cues (Brader 2006). 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).

These features of visual communication also highlight how the meaning of a visual argument does not inhere solely in the content of the communication itself. Rather, “the viewer’s interpretation of a visual argument is more a product of her or his own mind than it would be if the argument were completely explicit to begin with” (Messaris 1997, xviii). The crowd-sourcing technique we describe here allows scholars to capture multiple individual interpretations of ambiguous content.

The complexity of audiovisual communication makes ideal for political messages, but the lack of syntactic rules makes the measurement of political meaning—already a challenge in any medium—even more difficult for audiovisual material. If we are to advance understanding of visual political communication, we need more and better measurement of the content of actual communications. We will struggle to conduct cumulative empirical research that can advance theory development as long as coding video content is a time consuming, boutique enterprise. Drawing on the power of the crowd promises to make coding more efficient. As important, it can also make the coding process more open and therefore replicable and cumulative. This should facilitate empirical research that builds progressively and stimulates theoretical advances in our understanding of visual political communication in politics.

Video coding interface

We created a web-based coding portal to facilitate direct comparisons between coding populations. The RAs accessed it through our server; for the online workers it was embedded within the mTurk worker site (see figure 1). The interface showed coders a single ad to be coded, a data entry form, and an abbreviated summary of the coding instructions. Coders could expand the
instructions on-screen and could download the codebook as a PDF file. They viewed each ad one or more times while making their coding decisions. When workers submitted their coding, the system loaded the next randomly-selected ad. The portal also included a back-end interface for tracking progress, downloading data, and approving the mTurk work.

Figure 1: Coding Interface

The coding tasks and advertisements

To allow comprehensive comparisons between coding approaches, we included a diverse array of coding items. We sought concepts that (1) spanned the range from concrete/objective to abstract/subjective; (2) are communicated through a mixture of language and audiovisual elements; (3) varied in the political knowledge required to code; and (4) included items that replicate coding provided by the Wesleyan Media Project (WMP).

---

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

3 We accepted every submission except in two instances where workers consistently and repeatedly coded 30-second ads in less than 20 seconds each, suggesting that they were not actually watching them.

4 See online appendix A2 for additional information about the coding process.
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 further divide 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).

We included two examples of manifest content: the presence of the American flag and mention or appearance 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

<p>| Table 1: Coding items |</p>
<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>Does and American Flag appear?</td>
<td>yes/no</td>
<td>manifest</td>
</tr>
<tr>
<td>Candidate appearance in ad¹</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>Economic</td>
<td>Does ad include an economic appeal?</td>
<td>yes/no</td>
<td>pattern</td>
</tr>
<tr>
<td>Economic</td>
<td>Is economic appeal optimistic?²</td>
<td>yes/no</td>
<td>pattern</td>
</tr>
<tr>
<td>Emotional</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>Traits</td>
<td>Competence</td>
<td>no/ competent</td>
<td>latent</td>
</tr>
<tr>
<td>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>¹Appearance categories varied somewhat across waves; analysis based on collapsed, dichotomous version for “NO appearance” vs. any appearance \n²“FC” = favored candidate; “OC” = opposing candidate</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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. The full set of coding categories is detailed in Table 1.

These latent projective concepts present the biggest challenges for content analysis, because they are complex and symbolic, and therefore elude concise, rule-based definitions. Moreover, the audiovisual elements of campaign commercials are central to conveying emotions, traits, and perhaps even subtle ideological implications. Because of these features of projective concepts, they represent the sort of coding categories that are generally thought to require extensive face-to-face training to achieve reliable and valid coding.

Although these concepts are hard to define precisely and completely, they plausibly represent “primitive concepts”—that is, categories that most people understand intuitively (Potter and Levine-Donnerstein 1999, 260). This is clearly true for emotions. Most American coders, we expect, share understandings of fear, anger, and other emotions. With instructions that cue the right cognitive schema, we expect that coders should be able to identify and distinguish between them. Identifying traits may demand more specialized knowledge, though here too ordinary people should have relatively clear understandings of each. Coding ideological implication probably requires more specialized political knowledge, of course, but identifying the ideological implications in campaign ads would likely defy any attempt to develop a complete and objective set of rules for what counts as “liberal” or “conservative,” and so also relies on coders’ ability to recognize novel and creative expressions of ideology.

We deployed this coding system to measure the presence of these concepts in the universe of English-language campaign commercials aired in U.S. House of Representatives and Senate primary and general-election campaigns in 2010 (Fowler et al. 2014), which includes 4,357 unique

---

5 These trait categories followed guidelines developed by Hayes (2011) after scholars who find trait assessment central to candidate evaluation (Kinder et al. 1980; Holian and Prysby 2016).
advertisements (3,016 House and 1,341 Senate) produced by the campaigns and by political parties and other outside groups on behalf of a candidate. These ads represent 593 candidates in 279 races; collectively they were aired just over 1.5 million times.\(^6\)

**Traditional research assistant coders**

To replicate traditional content analysis, we hired undergraduate RAs whose recruitment, training, and working conditions followed standard practice.\(^7\) We emailed advertisements to political science majors and selected six students we knew to be diligent and 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. Then, between the meetings each RA coded a set of 30 practice ads. At the second meeting we discussed and resolved coding disagreements, further clarified the rules, and answered questions. Then the RAs began production coding; as they did so we stayed in touch to answer additional questions and clarify ambiguities.

**Online coding workforce**

We recruited online coders from Mechanical Turk (mTurk), an online labor market developed by Amazon.com to facilitate work that requires human intelligence. The mTurk system allows employers (called “requesters”) to create small, independent “human intelligence tasks” (HITs) for workers to complete. Workers can select from among thousands of available HITs that range from participation in academic research studies, to image classification, text translation, and much more.\(^8\)

Workers have incentives to do good work because requesters only pay for work they approve, and because Amazon makes workers’ overall approval rates available to requesters.\(^9\) Requesters can limit their HITs to workers with high approval percentages; they also can require that workers reside

---

\(^6\) There were 4.5 ads on average per House candidate (range 1-25; standard deviation 3.8), and 7.5 per Senate candidate (1-30; standard deviation 6.8).

\(^7\) The data collection was approved by the University of Virginia IRB, project number 2015031700.

\(^8\) mTurk was one of the earliest crowd source platforms; it has been used heavily to recruit social science research participants. Other vendors offer similar services, including crowd aggregators that provide access to multiple pools of workers (Vakharia and Lease 2015).

\(^9\) See online appendix A3 for additional details on requester reputation.
in the United States, or that workers pass a test or complete some task to earn a “custom qualification.” Workers also have a strong incentive to work quickly, while avoiding rejection, because they are paid by the task.

All contact between requesters and workers occurs through the mTurk website where tasks are posted, work is completed, and approval and payment arranged, or by email initiated by a requester or worker. This makes it impossible to conduct traditional, interactive face-to-face training and discussion. In theory one could pay workers to complete extensive online training. In practice, this presents a risk to workers whose work during training might not be approved, and a large challenge for requesters to design systems to enforce compliance and attention.

We seek not to reproduce this intensive training of a small number of RAs, but rather to evaluate an alternative: shorter, standardized training of a large number of lightly screened workers. We required workers to be U.S. residents, to have completed at least 100 HITs for any requester, and to have a 95 percent approval rate. We also required them to earn a custom qualification that verified they could see and hear the videos on their computer, collected background demographic and attitudinal information, and introduced them to the coding guidelines.

Overall, 1,076 mTurk workers took the qualification survey; all but seven were able to view the ads and were therefore classified as eligible for coding. Of these, 470 coded at least one ad. Each worker coded between one to 1,159 ads (median of 7; mean of 52; standard deviation of 121). The distribution is highly skewed: many dropped out after coding a few ads, moderate numbers coded dozens, and a few coding hundreds of ads. Seventy-five individuals completed 80 percent of the 19,089 ads coded by mTurk workers. Our mTurk workforce was reasonably diverse with respect to age, education, and income, though not representative of the American public as a whole. Twenty percent identified as Republicans and 45 percent as Democrats, and their political knowledge was

10 Hauser and Schwarz (2015) show that workers with high approval rates are generally quite attentive to the tasks they complete as subjects in academic research studies.

11 52 percent were women; the median age was 34 (range 18–81). 46 percent report having a college degree and 12 percent were students. The median coder reported income in the $40,000 to $80,000 range with 17 percent less than $20,000 and 19 percent above $80,000.
higher than the American public’s.\footnote{We included a standard political knowledge battery in our background survey that was modeled on the American National Election Study. Our median coder scored 0.875 on a zero-to-one scale.} About three quarters reported conducting at least some content coding in the past.

Both coding forces received the same coding manual and written instructions; beyond this their training differed. We knew the RAs and provided comprehensive training that included face-to-face discussion to tune their intuitions and refine and align their understanding of each coding category. The mTurk workers were unknown to us and less carefully selected. Their online, hands-on training occurred within a system incentivizing them to work quickly, but also carefully. Our online directions aimed to cue the right concepts and alert them to important distinctions, but we relied heavily on their intuitive understanding of concepts. The question is, how do the two workforces—one guided more by our training, the other drawing more on its priors—compare?

Comparing online and traditional coders

Our analysis focuses on the 20 coding decisions for each ad that we describe above and summarize in table 1. To facilitate comparisons, we purposely created overlap in the ads coded by our two workforces. Every ad was coded by at least five mTurk workers; a subset of 1,543 ads were also coded by one or more RAs. A group of about 200 ads was coded by at least 20 mTurk workers to allow more detailed reliability analyses; a further subset of 85 of those was coded by all six RAs. Each RA also coded 100 ads in common with one other RA and 188 unique ads. Across all types of coders, each ad was coded between five and 31 times (mean 6.9, standard deviation 4.1). Over the course of the project we adjusted coding instructions somewhat, so coders encountered some variation in instructions and coding items across several distinct waves of coding.\footnote{We collapse coding waves in our analyses, having found no evidence of systematic variations in reliability or validity.}

Reliability

Coding is reliable when different coders make the same decision about a piece of content. We measure reliability with Krippendorff’s alpha (Krippendorff 1970; Gwet 2014). This standard statistic 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, generalized for ordinal and interval data and for multiple non-unique raters. Happily, our results are unchanged when we substitute other reliability measures.

Table 2 presents summary reliability information for the different coding items, separately among RAs and mTurk workers, along with absolute and percentage differences in reliability between the two types of workers. The first column displays inter-rater agreement among the RAs; here we see substantial variation from item to item. Reliability is highest for whether the favored and opposing 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). 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. Concerned that disagreement on emotion coding may stem from the three-category measure (strong, weak, or no appeal), we created dichotomized versions in which each coding decision is collapsed to the presence (strong or weak) vs. absence of the emotion. Across these dichotomized versions, inter-coder agreement among the RAs is in fact the same, averaging 0.31 as well.

---

14 There are many alternatives for estimating inter-rater reliability and for weighting disagreements for non-binary coding; see Gwet (2014). Each statistic makes somewhat different assumptions about the rating process and each weighting differs in the relative penalty for smaller vs. larger disagreements when coding more than two categories. We use ordinal weights (Gwet 2014, eq. 3.5.1) for the three-category emotion coding and quadratic weights (Gwet 2014, eq. 3.2.5) for our continuous ideology coding. The results do not change with other weighting schemes. We use the Stata package kappaetc to calculate these statistics (Klein 2017).

15 Note that our interest focuses on relative reliability of different coder populations, rather than on the absolute reliability level. Online appendix A4 replicates our reliability analyses using Conger’s (1980) generalization of Cohen’s kappa.

16 See online appendix A4 for item-level reliability statistics.

17 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; for example, many include flag-like graphical elements that resemble or evoke the flag ambiguously.

18 This corresponds to down-weighting entirely disagreements between coding “strong” and “weak.”
Although some of these coefficients are low by conventional standards, they are in line with WMP’s own coding. Though we could not obtain exactly comparable reliability statistics from WMP for their coding of 2010 ads, they did provide kappa statistics for several comparable items in 2014, including manifest (flag) and projective (fear, enthusiasm, anger) items. Across these, the absolute levels of our RA reliability is comparable to theirs.\(^{19}\)

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

<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 emotions(^1)</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 \(\alpha\) 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 \texttt{kappaetc} (Klein 2017).

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

The second column of table 2 shows the corresponding inter-rater reliability statistics among mTurk workers. The individual mTurk workers look less reliable than traditional RAs. 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. These gains from aggregation might be a boon especially for coding subtle or ambiguous content apt to be missed by some individual coders, no matter how well trained.

\(^{19}\) Email communication, 2/17/2017.
Meta-coder reliability gains

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.”

To assess reliability gains from aggregation, we drew on the subset of ads coded by 20 workers. For each, we created a set four aggregates, constructed by averaging five workers randomly chosen from among the 20; we call each aggregate a “meta-coder,” because although it is based on multiple actual coders, we treat it as a single coding. To allow fully-comparable reliability analyses we rounded each average to produce the same set of outcome values (2, 3, or 101, depending on the item). (As we discuss below, we do not advocate this rounding in production content analyses, preferring to harness the information about the strength or clarity of the underlying stimulus that the unrounded average provides.)

Table 3 presents reliability gains from aggregation by examining agreement among these four meta-coders. The first column of the table summarizes the individual mTurk coder-level inter-rater reliability statistics; 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.

---

20 In online appendix A5, we show that hiring 4-5 coders is enough to decrease measurement error substantially.

21 In online appendix A4 we deploy a measure of reliability based on the mean-squared variation across coders to avoid rounding; the results are equivalent.

22 Benoit et al. compare simple averages of this sort with more complex Bayesian scaling models; they find that the two produce very similar results.
The fourth column shows agreement among the research assistants (for this subset of ads). The final two columns compare mTurk meta-coders with RAs, 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 RAs: meta-coder alpha was between 0.05 and 0.26 higher than RA coding, an improvement of five to 97 percent. For ideology, the meta-coders are equivalent to RAs; only on the presence of economic appeals do the meta-coders fall short of the RAs reliability, with alpha that is 0.14 lower.

Validity

Assessing validity is more difficult. 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. On this comparison, our results parallel those for reliability. Individual mTurk workers are less valid than a traditional RAs, but mTurk meta-coders achieve validity that equals and often greatly exceeds that of the RAs.

Table 3: Reliability gains due to meta-coders (Krippendorff’s α)

<table>
<thead>
<tr>
<th></th>
<th>mTurk workers</th>
<th>mTurk meta-coders</th>
<th>Difference: meta-coder gain</th>
<th>Research assistants</th>
<th>meta-mTurk vs. RA</th>
<th>meta-mTurk vs. RA (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average for candidate</td>
<td>0.86</td>
<td>0.97</td>
<td>+0.10</td>
<td>0.92</td>
<td>+0.05</td>
<td>+5%</td>
</tr>
<tr>
<td></td>
<td>(on meta</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>subset)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Economic appeal</td>
<td>0.34</td>
<td>0.39</td>
<td>+0.06</td>
<td>0.53</td>
<td>-0.14</td>
<td>-26%</td>
</tr>
<tr>
<td>Average for Economic</td>
<td>0.60</td>
<td>0.71</td>
<td>+0.10</td>
<td>0.66</td>
<td>+0.05</td>
<td>+7%</td>
</tr>
<tr>
<td></td>
<td>tone</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average for emotions</td>
<td>0.29</td>
<td>0.47</td>
<td>+0.18</td>
<td>0.28</td>
<td>+0.19</td>
<td>+67%</td>
</tr>
<tr>
<td>Average for dichotomized</td>
<td>0.31</td>
<td>0.33</td>
<td>+0.22</td>
<td>0.27</td>
<td>+0.26</td>
<td>+97%</td>
</tr>
<tr>
<td>emotions¹</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average for FC traits</td>
<td>0.32</td>
<td>0.56</td>
<td>+0.24</td>
<td>0.41</td>
<td>+0.15</td>
<td>+58%</td>
</tr>
<tr>
<td>Average for OC traits</td>
<td>0.33</td>
<td>0.49</td>
<td>+0.15</td>
<td>0.38</td>
<td>+0.11</td>
<td>+29%</td>
</tr>
<tr>
<td>Average for ideology</td>
<td>0.43</td>
<td>0.68</td>
<td>+0.24</td>
<td>0.66</td>
<td>+0.01</td>
<td>+2%</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).

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.
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. Of course, WMP’s coding is not completely valid itself. Nevertheless, it represents the best (and generally only) measurement we have of ad content. Moreover, these data are used widely to study political communication and its effects; as such, they represent an important validity standard.23

We calculate validity coefficients as the correlation between each criterion variable and the corresponding coding, separately among each workforce. We compare these validity coefficients for RAs, 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 RA 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 between the two correlations). However, the meta-coders do better, with a validity coefficient of 0.56—indistinguishable both substantively and statistically from RA validity.

We were able to devise 15 criterion variables that include 12 of our 20 coding categories. They are detailed in table 4, which also 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 RAs. 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

---

23 WMP data are used to study campaign negativity and voter mobilization (Goldstein and Freedman 2002a, 2002b; 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).
opponent lacks integrity (the criterion is WMP indicating that the ad mentions “corrupt” or “dishonest”). The fourth column shows the difference between RA 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.

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 worker validity</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*</td>
<td>474 - 1,492</td>
</tr>
<tr>
<td>PC appears (dichot)</td>
<td>WMP: FC mentioned or pictured</td>
<td>0.70</td>
<td>-0.01</td>
<td>1,485 - 1,510</td>
</tr>
<tr>
<td>OC appears (dichot)</td>
<td>WMP: OC mentioned or pictured</td>
<td>0.88</td>
<td>-0.02*</td>
<td>1,485 - 1,510</td>
</tr>
<tr>
<td>Economic appeal</td>
<td>WMP: Economic issue or mention</td>
<td>0.58</td>
<td>-0.18*</td>
<td>1,476 - 1,511</td>
</tr>
<tr>
<td>Emotion: enthusiasm</td>
<td>WMP: Enthusiasm appeal</td>
<td>0.41</td>
<td>-0.04</td>
<td>1,485 - 1,510</td>
</tr>
<tr>
<td>Emotion: enthusiasm</td>
<td>WMP: Uplifting music</td>
<td>0.42</td>
<td>-0.04</td>
<td>1,485 - 1,510</td>
</tr>
<tr>
<td>Emotion: fear</td>
<td>WMP: Fear appeal</td>
<td>0.38</td>
<td>-0.11*</td>
<td>1,485 - 1,510</td>
</tr>
<tr>
<td>Emotion: fear</td>
<td>WMP: Omnipresent tense music</td>
<td>0.24</td>
<td>+0.03</td>
<td>1,485 - 1,510</td>
</tr>
<tr>
<td>Emotion: anger</td>
<td>WMP: Anger appeal</td>
<td>0.35</td>
<td>-0.14*</td>
<td>1,485 - 1,510</td>
</tr>
<tr>
<td>Emotion: anger</td>
<td>WMP: Omnipresent tense music</td>
<td>0.45</td>
<td>-0.09*</td>
<td>1,485 - 1,510</td>
</tr>
<tr>
<td>PC strong leader</td>
<td>WMP: Ad mentions “tough,” “fighter,” “experienced”</td>
<td>0.32</td>
<td>-0.10*</td>
<td>1,302 - 1,338</td>
</tr>
<tr>
<td>FC integrity</td>
<td>WMP: Ad mentions “honest”</td>
<td>0.24</td>
<td>-0.15*</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*</td>
<td>897 - 1,049</td>
</tr>
<tr>
<td>PC ideology</td>
<td>FC DW-NOMINATE ideology score^2</td>
<td>0.58</td>
<td>-0.15*</td>
<td>465 - 495</td>
</tr>
<tr>
<td>OC ideology</td>
<td>OC DW-NOMINATE ideology score</td>
<td>0.62</td>
<td>-0.16*</td>
<td>200 - 237</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>0.48</td>
<td>-0.09*</td>
<td>465 - 495</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.

^2 First-dimension DW-NOMINATE score; available for candidates who held office at some point. Ideology analyses restricted to candidate-sponsored ads.

The next column shows that aggregation improves validity. The mTurk meta-coder validity almost always equals or exceeds—often substantially—that of the research assistants. As we saw above, coding of economic appeals by mTurk meta-coders is just as valid as the RAs’ coding. In 10 of 15 cases, the mTurk meta-coders produce higher validity than RAs; 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 less on attributions of integrity or lack thereof. Averaging across all 15, the meta-coders and RAs are essentially equally valid.

Particularly noteworthy are the final two rows of the table, which validate the ideology coding. For these items, coders were asked, “[b]ased on everything in the ad, where would you place the candidate ideologically? You should use your best judgment based on issue positions, traits, tone,
images, etc.” Judging ideological tone demands specialized political knowledge to identify projective latent content. To assess the validity of this difficult coding task, we turn to the external criterion of candidate ideology as revealed by voting behavior. For each candidate who served in Congress at some point, DW-NOMINATE provides an estimate of their ideological ideal point based on roll-call voting (Poole and Rosenthal 2007). In the final two rows of table 4, we validate the ideology coding against these ideology scores.²⁴ We observe the usual pattern: a strong positive correlation of ideology coding and DW-NOMINATE in the RA coding (0.58 for favored candidates; 0.62 for opponents), which implies strong coding validity. Validity is substantially lower among individual mTurk workers, by –0.15 and –0.16 for the favored and opposing candidates, respectively. Once we aggregate to meta-coders, however, the correlation is indistinguishable from that among the RA coding: 0.03 higher (n.s.) for the favored candidate and 0.05 lower (n.s.) for the opponent.

Resource comparisons

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

We paid our RAs $11 per hour (plus 6% fringe); this worked out to $0.56 per ad coding. In a full-scale project the costs would vary: on the one hand, training time would be amortized across more ads; on the other hand, we would double-code a subset of ads. Ultimately, RAs would likely cost between $0.50 and $0.60 per ad. Thus, the two workforces have comparable cost assuming we have each ad coded four or five online workers. The online workforce is substantially faster. At $0.10 per ad, we averaged one ad coding 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.

²⁴ This analysis includes only candidate-sponsored ads and eliminates those from parties and interest groups.
Additional advantages of multiple coding

We have stressed that aggregation reduces measurement error. But multiple coding brings other important benefits. Though most of our coding decisions are binary (or ternary), they are measures of continuous latent dimensions. For example, though we code for the presence or absence of the trait integrity, an ad can make more or less explicit, direct, and strong claims about a candidate’s honesty. Similarly, emotional appeals vary from nonexistent through overwhelmingly powerful; our ternary coding of absent, weak, and strong captures only some of this variation. Reliable coding of these items requires two things: that coders understand and agree on the nature of the concept—what counts as a claim about integrity, what counts as an appeal to anger—and also that they understand and agree on the threshold for counting it as in an ad. If some coders are more sensitive than others to the presence of a concept, then reliability will suffer even with complete agreement on a concept’s nature and indicators.

Moreover, working to align all the coders’ sensitivity thresholds can be statistically inefficient, as it discards information about the strength or clarity of a concept. When a single worker codes each ad, variation in sensitivity or coding thresholds reduces reliability. But when multiple workers analyze each item, variations in sensitivity can increase validity, by giving us multiple indicators of whether the concept actually appears, as well as information about its strength or clarity. As long as ads are randomly assigned to coders—as they were for our mTurk workforce—this generates an unbiased and more valid estimate of the underlying latent dimension.

Finally, collecting multiple measurements of each item from a large number of independent coders opens gives us leverage to model statistically the interpretive task surrounding that decision. 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. In order to focus as well on the individual coders, this could be extended to a non-nested model with coding decisions simultaneously grouped by ad and by coder (Gelman and Hill 2007, 241-248).
Transparency and replicability

Any scientifically respectable content analysis provides a codebook that documents the coding system. Nevertheless, reproducibility remains a major challenge. Developing coding guidelines, training coders, and implementing a traditional analysis all involve intensive interaction and discussion. These make the process unpredictable and therefore impossible to document completely. Though codebooks typically incorporate revisions that reflect modifications and clarifications to coding rules that arise, discussion and feedback are not so easily documented. Even with a written codebook we cannot fully replicate coder training because we cannot, even in principle, reproduce the same discussions.

The limits to online training and interaction become virtues in this context, because that training, though limited, is entirely standardized. Once the codebook is finalized and posted, it completely documents the coding and training system. This allows others to reproduce exactly that training with a new workforce. To be sure, the limited training and interaction with online coders almost certainly results in greater coder-level error variance, because we cannot work closely with coders to stamp it out through individual discussion, as we do under the traditional approach. However, our results show that we can mitigate this additional error variance through aggregation.

Moreover, an open and complete codebook of this sort facilitates “agile” data collection, in which we can modify coding systems for each project, and even experiment systematically with the coding guidelines. By randomly assigning training stimuli, researchers can study systematically the kinds of training materials that lead to the best coding, and can use empirical evidence to guide choices about procedures and to weigh trade-offs.

Discussion and Conclusion

Content analysis sounds deceptively straightforward. Researchers formulate a concept of interest and communicate it to coders, who search for indications of the content in the material to be analyzed. 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—which corresponds to with everyday understanding—holds that they 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 categories like these: for example, the manifest content coding category, “does an American flag appear in a campaign ad” 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). Objective rules can also delineate subtle boundary conditions (do flag lapel pins, Betsy Ross flags, or Jasper Johns paintings count?).

But many interesting categories—in life, in content analysis, in politics—are not defined by necessary and sufficient conditions. Wittgenstein’s concept of a “game” (1953) supplies an example that resonates with many political concepts. Games share no set of common features: for example, some involve luck, others skill, others both. Some have winners and losers, others do not. Moreover, the advent of video games expanded the concept (Lakoff 1987, 16). Categories like these are unified by “family resemblances,” not rules. Different games are linked as relatives are, by “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” incorporates 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 one of our projective categories, anger. Emotions “have an extremely complex conceptual structure;” and anger can expressed in a wide variety of ways (Lakoff 1987, 380). Lakoff spends 36 pages discussing the many synonyms, abstract metaphors, and prototypical scenarios that convey anger in order to explore fully the structure and limits of the concept; in this covers only the domain of explicit language, leaving aside audiovisual imagery (380-415).
It would be impossible to construct a truly comprehensive coding manual for the concept of anger. Moreover, because the audiovisual nature of campaign ads allows them to convey subtle, layered messages in symbolic, creative, and even ironic ways, they are likely to generate novel expressions of anger and of other complex constructs. For complex and ambiguous concepts like anger, intensive training of coders to “know it when they see it” (in Potter Stewart’s famous words about obscenity) supplements written rules. This is the traditional approach: by carefully and laboriously teaching the coders the precise nature and boundaries of each concept, we aim to instill in each an identical understanding of the concept, consistent with that of the subject-matter expert who developed the system. Specific examples plan an important role in this training, because in general, some examples of a concept are better, or more prototypical, than others (Rosch 1983). For example, robins and sparrows are excellent prototypes of the category “bird”; penguins and emus, though clearly birds, don’t feel as bird-like. We use specific examples help identify areas of conceptual complication or ambiguity—do penguin-like birds count, for the purposes of this analysis? As we train coders via discussions of better and worse category exemplars, we revise the coding manual to reflect these clarifications and revisions and to include specific examples. Through this process, we fine-tune the intuitions of the coders to reflect a nuanced understanding of the category that is shared among the coders and to create a document that might prompt similar intuitions and understandings by future coders.

However, as we note above, the complete coding system as implemented through the process of discussion and editing, cannot be fully documented. First, the discussion itself, along with the examples it engages, are implicitly part of the system. The coding guide might include some examples of what does or does not count, but it cannot include them all. More concerning, listing examples can cause as many problems as it solves, by focusing coders’ attention on particular instantiations of the concept, at the cost of other possibilities. More broadly, any written documentation of intensive face-to-face training will be incomplete. There is no way, even in theory, that a new investigator could reproduce with new coders the precise training and revision process itself.
It is well documented that different research teams can produce very different quantitative estimates of content. For example, Geer describes dramatic variation in coding of presidential campaign-ad negativity, with estimates varying from 24 percent to 54 percent (2006, 37). This is for a concept—negativity—that seems fairly straightforward on its face. Scholars hoping to replicate or extend any or all of these analyses could train research assistants using the respective coding guides. However, our contention is that they could never reproduce exactly the complete training experience, and therefore could not hope, even in principle, to produce in the minds of their coders a shared cognitive representation of the concepts being coded that exactly matches that of the original coders.

The open-codebook strategy for coding complex content we present forgoes this search for the holy grail of fully-aligned cognitive understanding among coders. Online workers who undergo limited but fully standardized trained online workforces make coding decisions with more measurement error, but we have shown that online efficiencies make up for individual error with increased quantity, allowing for gains from aggregation—as well as opening additional analytic possibilities. Rather than investing resources—both time and money—in stamping out variation among coders’ understanding of each concept, we instead embrace that variation, 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—what Potter and Levine-Donnerstein called “primitive concepts” (1999, 260). For example, we expect coders in the United States to have a good and reasonably well-shared sense of anger, and to be able to name it when they see it. Although they will vary somewhat in their implicit understanding of anger and in their sensitivity to it, with multiple measurements from different coders we can aggregate to a consensus judgment. Our findings suggest that this consensus judgement can be more reliable and valid than the individual judgement of intensively trained traditional RAs.

In closing, we offer four general recommendations for coding political video using online labor markets. First, have each video coded by multiple coders, which can produce comparable or
better reliability and validity than traditional RA coding. More importantly, repeated coding provides the researcher with an estimate of cross-coder variation at the level of each coding decision. This allows for the generation of continuous measures from easy-to-make binary coding decisions, and it generates opportunities for modelling aspects of the coding process itself.

Second, develop, pilot, and revise the coding system in-person, with a small team of experts or research assistants before deploying the production coding online. Nuanced feedback on new coding classifications is hard to collect online from coders who do not regularly interact with the researcher. By piloting instrumentation with RAs or other researchers, we can achieve some of the insights produced by intensive interaction before finalizing the standardized training instrument. We learned this the hard way. Initially we tried to separate sociotropic economic appeals, that refer to the overall economy, from pocketbook, appeals that touch on the viewer’s personal financial situation (Kinder and Kiewiet 1981). Reliability on these items was terrible; in training the research assistants and then reviewing many more ads, we discovered that actual campaign ads often mix these appeals in ways that makes them very hard to distinguish. Therefore, we dropped the distinction, as it did not appear meaningfully in the data. This experience suggests that scholars should not assume that theoretical concepts neatly map onto real world political communication, and underscores the importance of careful pilot testing of measures.

Third, focus on cueing the right general schema in the coding guide rather than developing exhaustive definitions for each concept. As we discuss above, long lists of criteria and examples are inevitably incomplete and online workers are unlikely to attend to them fully throughout their coding work. Here our advice diverges from the traditional approach, where we use in-person interaction to fine-tune coders’ intuitions. In the online context, focus instead on bringing the primitive concept to mind and on explaining any central and 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 also collects any relevant demographic or other background information. By restricting the coding task to those workers that
have demonstrated the ability to view and hear multimedia stimuli, researchers will avoid coders who
are not making a good faith effort to execute the task. The process of recruiting and training workers
also incentivizes continued coding work (as workers are invested in completing the initial survey)
and allows the researcher to revoke the qualification of those few workers that systematically ignore
instructions. We found that workers were extremely appreciative of our feedback, including bonus
payments, and some asked when new tasks would be made available.

Overall, then, this paper provides evidence that workers recruited from online labor markets
can classify political content—ranging from manifest to latent—with similar reliability, validity, and
cost as trained research assistants. What’s more, these workers save time relative to trained RAs,
while also making the entire coding sequence more transparent, replicable, and easier to extend and
adapt in future work. If multiple research teams undertake comparable coding projects using this
approach, their results will be easier to reconcile and both will benefit from the “miracle of
aggregation” we describe.
References


