COMMUNICATION METHODS AND MEASURES
2017, VOL. 11, NO. 3, 191–209
http://dx.doi.org/10.1080/19312458.2017.1317338

OPEN ACCESS

Content Analysis by the Crowd: Assessing the Usability of Crowdsourcing for Coding Latent Constructs

Fabienne Lind, Maria Gruber, and Hajo G. Boomgaarden

Department of Communications, University of Vienna, Vienna, Austria

ABSTRACT
Crowdsourcing platforms are commonly used for research in the humanities, social sciences and informatics, including the use of crowdworkers to annotate textual material or visuals. Utilizing two empirical studies, this article systematically assesses the potential of crowdcoding for less manifest contents of news texts, here focusing on political actor evaluations. Specifically, Study 1 compares the reliability and validity of crowdcoded data to that of manual content analyses; Study 2 proceeds to investigate the effects of material presentation, different types of coding instructions and answer option formats on data quality. We find that the performance of the crowd recommends crowdcoded data as a reliable and valid alternative to manually coded data, also for less manifest contents. While scale manipulations affected the results, minor modifications of the coding instructions or material presentation did not significantly influence data quality. In sum, crowdcoding appears a robust instrument to collect quantitative content data.

Crowdsourcing has become a major tool in business, arts and academia (e.g. Kittur et al., 2013; Shank, 2016). The use of crowdworkers as producers of quantitative data is increasingly prominent in social science or humanities research, informatics, and beyond (e.g., Hasegawa-Johnson, Cole, Jyothi, & Varshney, 2015; Shank, 2016). Taking advantage of the "wisdom of the crowd" (Surowiecki, 2005), crowdsourcing platforms are used by scholars from various disciplines to generate various sorts of data (Saur-Amaral, 2012; Saxton, Oh, & Kishore, 2013). Howe (2006) coined the term *crowdsourcing* in his article and refers to it on the entry page of his blog (http://crowdsourcing.typepad.com) as "the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call".

In communication and political science, particularly survey and experimental research has benefited from crowdsourcing as an efficient tool to recruit research participants (e.g., Berinsky, Huber, & Lenz, 2012; Peer, Samat, Brandimarte, & Acquisti, 2016). A range of studies is concerned with the implications and validity of experimental (or survey) data generated through the use of crowdworkers (Behrend, Sharek, Meade & Wiebe, 2011; Hill, Dean, & Murphy, 2014; Komarov, Reinecke, & Gajos, 2013; Levay, Freese, & Druckman, 2016; Snow, O'Connor, Jurafsky, & Ng, 2008; Wang, Huang, Yao, & Chan, 2015). For experiments in political science it appears that the advantages of crowdsourcing may outweigh its potential downsides (e.g., Mullinix, Leeper, Druckman, & Freese, 2015). Another common application is to employ crowdsourcing for coding jobs thus crowdcoding to annotate texts, visuals or audiovisual materials (e.g., Oosterman, Yang, Bozzon, Aroyo, & Houben, 2015; Wang et al., 2015). In particular, the annotation of social media contents has gained prominence over the past years (e.g., Simpson et al., 2015).

The use of crowdsourcing for quantitative content analysis in (political) communication research so far is less common (for exceptions see Budak, Goel, & Rao, 2016; Haselmayer & Jenny, 2014, 2016).

CONTACT Hajo G. Boomgaarden hajo.boomgaarden@univie.ac.at Department of Communications, University of Vienna, Rathausstraße 19, 1090 Vienna, Austria.

Methodological innovation in content analysis procedures rather focuses on automated, computer-assisted techniques (e.g., Boumans & Trilling, 2016; Grimmer & Stewart, 2013; Jacobi, van Atteveldt, & Welbers, 2016). While traditional, manual content analysis is described as a costly and labor intensive methodology (Krippendorff, 2013), the precision and validity with which computers implement content analysis in its various forms is still in development and often depending on the input of human coders in the form of training data for dictionary or algorithm development (e.g. Burscher, 2016). Given the persistent centrality of human coders for content analysis in communication research, this article assesses the adoption and handling of crowdsourcing as an efficient means to conduct quantitative content analysis. It does so by focusing on the analysis of rather latent, or as it can be also referred to, on a less manifest type of content characteristic (Neuendorf, 2002)—the evaluation of political actors. Potter and Levine-Donnerstein (1999) distinguish pattern and projective latency of content features. While the former focuses on patterns in the content itself, the latter is concerned with coders' interpretation of the actual content. It is this projective latency of content, by others referred to as subjective (e.g. Haselmayer & Jenny, 2014) or interpretative coding, that we take into account here. Due to the digital availability of various sorts of texts and increasing computer power to analyze large volumes of texts, great advances have been made in terms of automated, computer-assisted content analysis (e.g., Boumans & Trilling, 2016; Krippendorff, 2013). A common adoption relates to content that clearly manifests itself in the use of pre-defined words and phrases (i.e., dictionaries) (e.g., Pang & Lee, 2008; Riloff & Wiebe, 2003). When it comes to less manifest types of contents, manual content analysis remains a standard, either in the form of stand-alone analysis or to provide training sets for machine learning procedures. We argue that crowdsourcing is a very efficient and cost-effective tool for the production of quantitative content data, in particular with regards to contents that are harder to pre-define in dictionaries, i.e., less manifest, projectively latent constructs.

Using crowdsourcing for such type of content analysis, however, has seemingly clear disadvantages compared to classic manual analysis: Coders cannot be pre-selected based on experience, long term commitment of coders to the project is not assured, there is no opportunity for coder training in a traditional sense, and extended coding instructions are rather uncommon. But these disadvantages are met by the clear advantages of crowdsourcing. The data generation process can be potentially very quick, even for larger amounts of data, and it likely comes at considerably lower costs than a traditional manual approach. Furthermore, crowdcoded content analysis data may potentially be more reliable and easier to replicate (Benoit, Conway, Lauderdale, Laver, & Mikhaylov, 2016). While the discipline has rather standardized procedures for manual content analysis, such are lacking for crowdsourced content analysis.

The aim of this article is twofold: Study 1 considers the reliability and validity of the data generated by the crowd. If content analysis of latent constructs shall be outsourced to crowds, we need to have some understanding of the reliability with which crowdworkers generate content data and of how valid these data are, and how to treat the data to yield higher levels of validity. To that end we compare data from identical materials in different aggregations of a crowdcoding procedure to manually coded data. In Study 2 we relate to concrete working routines to be applied when using crowdsourcing tools for content analysis. Based on a systematic, experimental comparison of different ways of presenting texts to be coded by the crowd, and of different types of coding instructions and answer option formats, this section works towards a best-practice type of standardization of such coding procedures. In both studies we focus on the measurement of the rather latent content characteristics of political actor evaluations in news texts. More specifically here, the coding tasks are basically sentiment analyses that aim at identifying evaluative tendencies regarding particular target objects on the sentence level, which can be represented by classifying judgments as positive, negative or neutral (Mohammad, 2016a). Sentiment analysis refers to the task of determining the polarity or valence of words, sentences or documents.

In sum, both studies should be highly informative for scholars intending to use crowdsourcing for content analytical purposes. They are relevant and innovative in that they (1) systematically engage with crowdsourcing as a methodological tool for quantitative content analysis, (2) in an encompassing manner assess the reliability and validity of crowdsourced data by comparing it to manually coded data, and (3) systematically evaluate the consequences of variations in how coding materials, coding instructions, and answer option formats are presented to the crowdworkers for data quality. The studies are implemented in the crowdsourcing platform CrowdFlower (CrowdFlower.com). While Amazon's Mechanical Turk appears to be the currently most cited platform in the social sciences, CrowdFlower is generally described as equally appropriate (Vakharia & Lease, 2015). We are confident that our assessments will spark a more reflective and wider application of crowdsourcing for quantitative content analysis.

Crowdsourcing for content analysis

Social science research has increasingly relied on the efforts of crowdworkers lately, especially in terms of data collections (Shank, 2016). Despite the anonymous setting, crowdworkers genuinely care about their performance quality and are not exclusively motivated by financial incentives (Mason & Suri, 2012). Moreover, crowdsourcing samples do not differ substantially from population-based samples, which make crowdsourcing platforms a promising alternative to common convenient sampling for conducting experiments (Levay et al., 2016).

Crowdsourcing platforms are also particularly well suited for tasks that include translating or annotating text as well as video and audio materials (Shank, 2016). One example is the automated detection of sentiments in textual material by applying different models of classifiers, algorithms and combining crowdsourcing approaches with machine learning approaches (e.g., Budak et al., 2016; Hsueh, Melville, & Sindhwani, 2009; Simpson et al., 2015). Obvious advantages relate to the efficiency with which crowds can be used for such annotations, both in terms of time and budget. Apart from efficiency, concerning quantitative content analysis and its coding procedures, some scholars argue for the use of crowdsourced data instead of obtaining data coded by trained experts for reasons of reliability as well as reproducibility. Benoit et al. (2016, p. 278) state, that "empirical social science often relies on data that (…) are transformed into quantitative variables by expert researchers who analyze and interpret qualitative raw sources. (…) this expert-driven process is inherently difficult to replicate or to assess on grounds of reliability." The authors bring forward the argument that comparable results can be obtained through the performance of numerous non-experts in a quicker and cheaper manner, including the great benefit of being "intrinsically reproducible" (Benoit et al., 2016, p. 278).

Selection of workers, reliability, and validity

Despite the advantages, there are also some obvious downsides in using crowdworkers for content analysis. These largely concern the lack of control over who is working and the lack of training opportunities. Hence, when taking untrained crowdworkers as a substitute for manual coders, some concerns regarding workers' attitudes, the handling and evaluation of the task and thus the *reliability* and *validity* of the data may arise. Generally, it is difficult to determine the quality of a contributor's work on crowdsourcing platforms. Vuurens, De Vries, and Eickhoff (2011) differentiate between *ethical workers*, who accurately read given instructions and are willing to produce useful results, and *spammers*, who aim at earning the greatest possible amount of money by accomplishing jobs in the fastest and easiest way. Taking the existence of spammers into account, one needs to be aware of the possibility of a crowd providing a high amount of noise instead of meaningful results. Following established procedures in content analysis, intercoder reliability (ICR) needs to be assessed, as otherwise "the data and interpretations of the data can never be considered valid" (Lombard, Snyder-Duch, & Bracken, 2002, p. 589). To demonstrate the "trustworthiness of their data" (Krippendorff, 2013, p. 268) researchers need to consider to what degree their data would be replicable in a different context. Coders provide inter-subjective judgements, which can be assessed by ICR measures, such as Krippendorff's alpha which is preferred here given its flexibility and strictness (Krippendorff, 2013; Lombard et al., 2002). Assessing ICR in crowdcoding is easy and challenging at the same time. On the one hand, without too much effort and resources identical materials can be coded by several workers more or less simultaneously, thereby generating data appropriate for ICR assessments. On the other hand, over longer periods it is unlikely that the same workers would contribute to the task throughout, which in turn means that ICR is estimated over the contributions of different sets of coders. Furthermore, given that crowdsourcing procedures are less interested in the judgment of a single worker, but rather in the performance of the crowd—and thus would usually have multiple workers performing the same task on identical material (see section below)—ICR can be used to not only assess the quality of the task, but also to consider the performance of the crowd for the entire task and data reliability for different aggregation procedures.

Determining less manifest contents, such as political actor evaluations, depicts a complex task—even for human coders—since such evaluations are not necessarily bound to manifest content characteristics that can always be pre-defined. As a result of the subjectivity that comes with such judgments and evaluations, it is much harder to reach conclusive agreements among different coders than in other coding tasks. This poses challenges to data reliability, which, however, do not automatically translate into low validity of the data. In terms of coding of less manifest contents, obtained through crowdsourcing platforms, a group decision may reveal a more valid result than the judgment of one single person (Haselmayer & Jenny, 2014, 2016). After all, "under the right circumstances, groups are remarkably intelligent, and are often smarter than the smartest people in them" (Surowiecki, 2005, p. XIII). Thus, when considering subjective rating tasks, Liu (2012) states, that it is in fact essential to gather a collection of opinions from a number of different people. Consequently, for content analysis tasks multiple workers would provide judgments on the same material throughout the entire task. While a greater number of coders would naturally increase the chances for greater variation in the data, which would lead to somewhat lower IRC scores than usual (e.g., Haselmayer & Jenny, 2014), it is still likely that aggregate scores would project valid answers.

Acknowledging the general importance of various facets of validity (such as face or social validity), considering the aims of Study 1 we here focus on empirical validity, "the degree to which specific inferences withstand the challenges of additional data, of the findings of other research efforts" (Krippendorff, 2013, p. 331), also referred to as criterion validity (Neuendorf, 2002, p. 115). More specifically, we estimate concurrent validity (Krippendorff, 2013, p. 334), in that we compare the findings obtained by the crowdsourcing procedure with findings from more established types of content analysis.

In terms of minimizing the risks of low reliability and validity, crowdsourcing customers can imply certain quality mechanisms within the platform regarding the data collection process on the one hand and on the other hand concerning the handling of the data. *First*, so-called *gold sets* or *test questions* depict representative questions of a given crowdsourcing task, to which the correct answers are already known. This kind of questions is used to evaluate the quality of crowdworkers' performance. When they fail too many test questions, they are automatically removed from the pool of workers. As the concept of test questions is easy to understand and its implementation is simple, it is a common practice to filter out spammers (Vuurens et al., 2011).

Second, in the settings option of most crowdsourcing platforms, customers are enabled to select the performance level of their crowdworkers. This performance level is determined by the accuracy of crowdworkers' answers to more than a hundred test questions with predetermined correct answers across a large set of different jobs. A higher performance level indicates a higher accuracy. Although it might take longer to finish a job when only accepting high quality workers it depicts another convenient opportunity for spam-reduction.

Provided that data about the quality of workers is available, researchers have focused on establishing the most favorable way of aggregating data gathered through crowdsourcing. One of the simplest approaches is the technique of majority vote, for which the data are aggregated in such a way, that the answer mentioned most frequently illustrates the 'correct' one, the one that is selected for further processing of the data. Although this method returns meaningful results (Hsueh et al., 2009), it has also been criticized as it assumes all crowdworkers to be equally reliable, while it is evident that the quality and performance level differs within the group of workers (Vuurens et al., 2011). Other approaches take the differences in crowdworkers' skill levels (either their general performance level or their performance regarding the test questions of the current job) into account when aggregating the data (Simpson et al., 2015). On request crowdsourcing platforms also provide customers with aggregated results. On CrowdFlower, for instance, the aggregated results file includes the answers with the highest reliability, based on individual workers' trust ratings, so-called trust scores. These stem from the accuracy of crowdworkers' answers to the given test questions of the specific job. From the collected data pool, customers can decide freely how many trustworthy annotations per unit they want to receive. Moreover, CrowdFlower provides the confidence level of each score, which identifies the inter-annotator agreement (similar to ICR) and is a measure of CrowdFlower's confidence in the validity of the results.[1]

Study 1 assesses both the ICR and the empirical, concurrent validity of the results of crowdcoded content analysis data. It estimates Krippendorff's alpha and compares the scores of different aggregations of the crowd data and manually coded data to speak to the appropriateness of crowdcoded data in terms of ICR. With regards to validity, it compares the outcomes of the crowd procedure, again on different aggregation levels, with the outcome of manually coded data. It thereby establishes whether crowdcoded data in general, and in particular which types of aggregation measures produce more reliable and valid results in the case of coding less manifest, more subjective types of contents.

Material presentation, coding instructions, and answer option formats

Another way to reduce noise in crowdcoded data is an accurate and sophisticated design of the job, equivalent to the importance of high quality coding instructions and training in traditional manual content analysis. Prior research deals with the effects of task design on crowdworkers' performance (e.g., Catallo, 2015; Finnerty, Kucherbaev, Tranquillini, & Convertino, 2013; Kazai, Kamps, Koolen, & Milic-Frayling, 2011). When we created the coding task for Study 1, many decisions were based on their findings (e.g., appropriate payment, job title design, etc.). However, given the early stage of the literature about using crowdsouring platforms for content analysis research, there is little in terms of recommendations as to how to implement content analysis tasks in a crowdsourcing platform. Based on prior literature, however, it is reasonable to assume that minor modifications within the design settings of content analysis tasks can influence coders. Study 2 therefore focuses on design issues related to the presentation of the coding material, the instructions given to the workers, and variations in answer option, three factors the literature is rather silent about. Doing so allows us to follow a broader goal, namely to work toward best-practice recommendations regarding the most appropriate design of a crowdsourcing task to code latent constructs in news texts.

Material presentation

According to Mohammad (2016b) the task of determining the target of opinion in a sentence can be challenging for coders and may lead to confusion. Crucial are the target specification and degree of anonymity (Mohammad, 2016b). For our study, this begs the question of how one can make sure that the crowdworkers correctly identify the target of opinion—political actors or parties—within a sentence, and are unaffected in their judgment by their personal attitude towards them. Anonymization and marking of targets in the material to be coded are obvious, but labor-intensive procedures and we have no empirical evidence whether such investment is necessary for content analysis with crowdworkers.

Coding instructions

To obtain annotations of high quality, instructions need to be simple, clear, and free from any ambiguity (Mohammad, 2016a), a recommendation which is not as easy to implement as it seems at the first glance. For example, by reviewing the codebooks of political actor evaluations in traditional content analyses, greater variation in the wording of coding instructions was evident. Some ask for "explicit evaluations", some just for "evaluations", some to "take the perspective of the target of opinion" (e.g. Banducci, de Vreese,

[1] If the data are already gathered, "bad" responses can also be filtered out through an examination of response patterns (Zhu & Carterette, 2010) and the duration of the task (Kittur, Chi, & Suh, 2008). We analyzed response patterns and used a CrowdFlower setting tool that automatically refuses the work of contributors that fall below a minimum time specification for a task.

Semetko, Boomgarden, & Luhiste, 2010; Schönbach et al., 2016; Weiß, Maurer, Schwotzer, Trautmann, & Zhu, 2009). A template of the CrowdFlower crowdsourcing platform recommends to include the sentence "Please do not overuse neutral-no evaluation" in the instruction. Again there is little if anything in terms of general guidelines how to phrase the necessarily short, yet concise instructions in a crowdworking environment.

Answer option formats

Finally, examination of the effects of different answer option formats, in particular those studied in the context of questionnaire design research (e.g., Bradburn & Sudman, 1979) and more specifically for sentiment measurements (Heise, 2010) were considered. Among the factors known to influence responses are the order in which the answer options are presented (e.g. primacy effect), the range of alternatives, and the order of questions (Schwarz, Knäuper, Oyserman, & Stich, 2008). Transferring such long studied effects to a new setting, we consider how crowdworkers' annotations are affected by variations in answer option formats.

In order to provide systematic evidence for formulating best-practice recommendations for the coding of evaluations in texts, Study 2 draws on an experimental design in which we manipulate the material presented, the coding instructions, and answer option formats. We explicitly acknowledge the exploratory nature of this endeavor. Given that we have virtually no evidence that would allow formulating specific expectations regarding the appropriateness of the three aspects in the job design for crowdsourcing platforms, we refrain from doing so here.

Study 1: Validation of crowdcoded data

Data and methods

By looking at sentiments towards political actors in news texts, the crowdsourcing task proposed in this study is of interpretative nature, relating to latent content features and thus can be described as rather subjective (see above). To assess whether our crowdcoded data is valid, we compared it to data from manual coders' judgments, equating empirical concurrent validation procedures (Krippendorff, 2013). Specifically, we draw on two independent sources of reference. On the one hand, five communication science graduate students (in the following referred to as "offline coders") were inquired to perform the same task as the crowdworkers—but offline and within a controlled setting. On the other hand, and as the third data source, we used already available material and data from the media content analysis within the Austrian National Election Studies (AUTNES) 2013 (Eberl et al., 2016). Here, seven coders (in the following referred to as "AUTNES coders"), assessed the sentences after attending a coder training of several days. Table 1 provides an overview of the three different data sources, the crowdworkers, the offline coders, and the AUTNES coders.

Material sample

The task performed by all coders, the crowdworkers, the offline coders, and the AUTNES coders included the judgment of latent content, a political actor evaluation, on the sentence level. The sentence sample for the crowdworkers and offline coders consisted of the same set of 500 German-language sentences from Austrian political news coverage. 250 of these sentences are a random sample from the sentences that were coded for the AUTNES media study (Kleinen-von Königslöw et al., 2016). This AUTNES data set contains more than 50,000 sentences from political news coverage during the Austrian national elections campaign in 2013.

The 250 other sentences, that we sampled additionally, focus on the coverage of political news in the time after the Austrian national elections in 2013, starting with the day after the nomination of the new government (December 17, 2013) until February 2016. For reasons of comparability, we followed the sampling approach of the AUTNES study, comprehensively described in the project documentation (Kleinen-von Königslöw et al., 2016). Hence, we also used the database of the Austrian Press Agency,

Table 1. Overview of the data sources crowdworkers, offline coders and AUTNES coders [study 1].

Data source	Number of coders	Number of sentences	Judgments per sentence	Setting	Training
Crowdworker	158	500	10	Online, no control [a]	Indirect via test questions, short introduction [b]
Offline coders	5	500	5	Offline, control	No training, short introduction [b]
AUTNES coders	7	250	1	Offline, control	Task specific training, comprehensive instruction

Note. For more details on AUTNES please refer to Kleinen-von Königslöw et al. (2016).

[a] "No control" means that as the coding was done online, we have no information about the diligence or surrounding of the crowdworkers while they were completing the task.

[b] The short introduction included explanations on the two-step evaluation task as well as on the fact that coders should focus on the evaluation of the actor marked with X. It was identical for the crowdworkers and the offline coders and is available on request.

the same newspaper sources (Der Standard, Die Presse, Kronen Zeitung, Salzburger Nachrichten, Österreich, Heute, Kurier, Kleine Zeitung), an updated version of the AUTNES search string (as holders of political offices changed after the national elections) to draw a random sample of news items within the period of investigation, and the same procedure for the selection of sentences from the retrieved articles. The search string contained names of the political parties in Austria, names of ministers and other politicians and is available on request. All names of political actors or political parties were replaced with the phrase "Political actor" or "Party" and we clearly defined the target of opinion in every sentence by marking them with the label "Political Actor_X" or "Party_X". Hence, crowdworkers and offline coders dealt with sentences such as "Imagine a world without POLITICAL ACTOR_X" or "But POLITICAL ACTOR_X intends to take care for the foster children".

The crowdsourcing task

The coding of the 500 sentences was conducted via CrowdFlower. On Mturk, another frequently used crowdsourcing platform, only customers from specific countries are allowed to launch jobs, whereas there are no such restrictions on CrowdFlower. Furthermore, we chose CrowdFlower as it provides a high number of German-speaking contributors.

First, crowdworkers were asked if the sentence would contain a rating of the target of opinion (Political Actor_X or Party_X) and if so, in a second step they were asked how the target of opinion is rated based on a five-item scale consisting of the response options "explicitly negative", "rather negative", "mixed (both positive and negative)", "rather positive", and "explicitly positive".[2] While the AUTNES coding was based on a three-item-scale consisting of the labels "approval", "neutral", and "criticism",[3] the crowdsourcing coders were exposed to two steps of evaluating the sentiment. Additionally, test questions ($N = 15$) in form of representative sentences of the given task were created, to which we had previously defined correct answers. Given that our rating task aimed at evaluating latent content based on subjective interpretations of contributors, we accepted multiple answer options for each test sentence, including those that were at the edge of the scale range we could fully agree upon within the research team. Thus, to give an example, the crowdworkers received a positive feedback when they rated a test sentence that contained a negative actor evaluation according to the understanding of the three researchers, with "explicitly negative", "rather negative", or "mixed (both positive and negative)". These test sentences originate either from the AUTNES sentence database or were collected with the exact same approach as the 250 additionally sampled sentences. They kept track of workers' performance quality and were randomly presented to crowdworkers throughout the job. The crowdworkers could therefore not distinguish if they were just about to code a test sentence or a sentence that is part of the job. But they

[2] Although it would have been preferable to use coding instructions identical to the AUTNES study, we were not able to implement the AUTNES coding procedure into the crowdsourcing task's design considering its length and complexity and crowdworkers' possible unfamiliarity with the process of content analysis. Therefore, instructions that differed from the AUTNES-original regarding the fixation of the target of opinion and ease of language were used to guide the online contributors.

[3] Variable V31 (Schönbach et al., 2016).

received immediate feedback and the correct answer when they failed a test sentence, a procedure that we classify as on-the-job training.

We ordered ten judgments for each sentence and received 5,059 judgments by 158 different crowdworkers within 6 days and 14 hr, including by default each worker's trust score which is assessed by their performance in answering the test questions. On CrowdFlower, the workers whose answers to test questions match with the predefined answer to a high degree receive a high trust score, whereas bad performance in answering test questions results in lower trust scores. The work of an online contributor is fully excluded when he or she has failed too many test questions.

Only German native speakers from Austria or Germany that were rated as highest quality workers (level three out of three) by CrowdFlower were given access to the job. In sum, 158 crowdworkers, 23% from Austria and 77% from Germany, coded on average 32 sentences, with a maximum of 40 sentences per worker and a minimum of 3 sentences. Crowdworkers were paid $0.15 per judgment of five sentences. More than half of the crowdworkers ($N = 80$) gave feedback to our task via a standard survey that CrowdFlower asks them to fill out after finishing a job. Out of 5 possible points for each category, the payment was rated with 3.9, the clearness of instructions with 4.2, fairness of test questions with 3.9, and ease of job with 3.7 points. Overall crowdworkers' satisfaction with the task reached therefore a score of 4 out of 5 possible points.

Offline coding task

As mentioned above, our coding instructions for the crowdworkers differed from those used in the AUTNES study. Therefore, we recruited five additional offline coders that were confronted with the exact same coding instruction and scales as the crowdworkers. Each of them coded all 500 sentences. In contrast to the crowdsourcing approach, here we were able to ensure the qualification and suitability of the offline coders as they were selected from a pool of Communication Science students, familiar with content analysis and coding procedures. Moreover, in this setting the authors could control for the environment in which the five offline coders, aged between 23 and 29, one male and four females, conducted the coding task.

Results

Intercoder reliability

To assess reliability, we calculated Krippendorff's alpha values for the data coded by the crowdworkers and offline coders and contrasted them with those reported in the study documentation of AUTNES (Kleinen-von Königslöw et al., 2016).[4] The subjectivity of the task of this study needs to be kept in mind when interpreting the results of the ICR measures (Haselmayer & Jenny, 2014; Neuendorf, 2002). Comparing the alpha values of the crowdworkers, of offline coders and AUTNES coders, it becomes evident, that they do not differ much. We consider the Krippendorff"s alpha values depicted in Table 2 as satisfactory, but far from perfect values regarding the actual sentiment coded.

In general, values below .8 indicate low agreement (Krippendorff, 2013; Neuendorf, 2002). A point worth considering here is the fact that compared to regular content analyses conducted mainly by just a few coders, the variety of crowdsourcing workers is much higher. For every sentence, judgments of ten coders were used. However, the sentences were coded by different groups of ten individual workers; there were not even two sentences that were coded by the same set of coders. Furthermore, we stress that the crowdworkers and the offline coders had not received a coding training in a "traditional" sense. Following that, we acknowledge that in terms of common standards for traditional content analysis the ICR values do not suffice, but in this study's context we accept them as not optimal but tolerable.

[4] ICR assessment for the AUTNES data is those reported for variable V31 "object evaluation" in the AUTNES documentation (Kleinen-von Königslöw et al., 2016). For AUTNES overall more than 50,000 sentences were coded by seven coders, the ICR measures were calculated based on their coding of 790 sentences.

Table 2. Intercoder reliability of crowdworkers, offline coders and AUTNES coders (Krippendorff's alpha) [study 1].

Data source	Evaluation dichotomous	Evaluation tendency 5-point scale	Evaluation tendency 3-point scale
Crowdworkers [a]	0.27	0.60	
Crowdworkers Top 5 [b]	0.31	0.66	
Crowdworkers Top 3 [c]	0.34	0.66	
Offline coders [d]	0.38	0.73	
AUTNES coders [e]			0.76

[a] Coders per sentence = 10; in total 158 different coders contributed; each of them coded up to 40 different sentences; hence, the coders were not identical for each sentence, which depicts an uncommon setup for ICR measures.

[b] Coders per sentence = 5; in total 116 different coders contributed; the Crowdworkers Top 5 is a subsample of the full crowdcoded sample, here we selected the five 'best' answers (those coded by the workers with the highest trust scores) per sentence.

[c] Coders per sentence = 3; in total 97 different coders contributed; the Crowdworkers Top 3 is a subsample of the full crowdcoded sample, here we selected the three 'best' answers (those coded by the workers with the highest trust scores) per sentence.

[d] Coders per sentence = 5; five different coders contributed, each coder coded all 500 sentences.

[e] Coders per sentence = 7; reliability test procedure and results for variable V31 "object evaluation" reported in the AUTNES documentation (Kleinen-von Königslöw et al., 2016). For AUTNES overall more than 50000 sentences were coded by seven coders, the ICR measures were calculated based on their coding of 790 sentences.

Interestingly, our coders appear to disagree more on whether the sentence contains a rating of the target of opinion, than on how the target of opinion is evaluated (see discussion). When comparing Krippendorff's alpha values of the whole crowdsourcing sample (ten judgments per sentence) with those of the five/three best rated judgments (according to crowdworkers' trust scores, see notes to Table 2) the values increase slightly, however they did not quite reach the ICR level of the offline coders (also see discussion below).

Ratings

Concerning the validation of the crowdsourced data, we contrasted all ratings by crowdworkers (N = 5059) with the annotations of the offline coders (N = 2489) on the one hand and subsamples of these data sources with the AUTNES coding (N = 196) on the other hand. We discuss differences and similarities between the groups in terms of the actual aggregated outcomes with all data shown in Table 3 below (for information on sample sizes see notes to Table 3).

Overall with regard to the first item, offline coders selected "evaluation" less often (62%) than "neutral—no evaluation" (38%) compared to the crowdworkers (evaluation = 74%; neutral-no evaluation = 26%). This statistically significant difference suggests that crowdworkers more often perceive an evaluation of the target of opinion compared to the offline coders, $\chi 2(1, N = 5748) = 121.49, p < .001$. When we compared the two groups with regard to their judgments of evaluation tendencies (5-point scale ranging from -2 = explicitly negative to 2 = explicitly positive), it shows that offline coders' annotations are on average slightly more negative ($M = -0.60, SD = 1.28, N = 1533$) than those of the crowdworkers ($M = -0.30, SD = 1.21, N = 3744$), a statistically significant difference, $t(2709) = -7.77, p < .001$.

We obtained highly similar results when we weighed the crowdworkers' results according to their trust scores.[5] Again, the weighed crowdworkers' annotations are significantly different compared to those of the offline coders, both with regard to the first decision "evaluation" vs. "neutral-no evaluation" ($\chi 2[1, N = 7027] = 117.55, p < .001$) and to the annotation of the evaluation tendency, $t(2826) = -7.73$, $p < .001$. But they are virtually identical in terms of central tendencies compared to the sample not taking into account trust scores ("unweighted data").

For 196 out of the 250 sentences that were collected from the AUTNES data base we were able to compare our data to the AUTNES coding.[6] For this purpose, the ratings of the crowdworkers and offline coders were recoded to the 3-point scale (-1 = criticism, 0 = neutral, 1 = approval) used by the AUTNES coders. The effect of the data source on the rating results was found to be significant, $F(2, 539) = 8.32$,

[5] The crowdworkers' trust scores ranged from 0.73 to 1 (M = 0.90; SD = 0.07), all offline coders received a trust score of 1. As a consequence, 4,538 annotations (Crowdworkers weighed with trust scores) instead of 5,059 annotations (Crowdworkers unweighted) were taken into account for the chi-square test and 3,361 annotations (Crowdworkers weighted with trust scores) instead of 3,744 annotations (Crowdworkers unweighted) for the t-test.

[6] In 54 cases we had to dismiss the AUTNES results as source of comparison, as the rating differed in terms of the target of opinion.

Table 3. Ratings of crowdworkers, offline coders, and AUTNES coders [study 1].

	Evaluation dichotomous [a]		Evaluation tendency 5-point scale [b]		Evaluation tendency 3-point scale [c]	
Date source	Mean (*SD*)	*N* [d]	Mean (*SD*)	*N* [e]	Mean (*SD*)	*N* [f]
Crowdworkers	0.74 (0.44)	5059	−0.30 (1.21)	3744	−0.20 (0.77)	1981
Crowdworkers weighted according to their trust score	0.74 (0.44)	4538	−0.30 (1.21)	3361	−0.20 (0.77)	1768
Offline coders	0.62 (0.49)	2489	−0.60 (1.28)	1533	−0.31 (0.74)	979
AUTNES coders					−0.14 (0.69)	196

[a] Coded as: 0 = neutral-no evaluation, 1 = evaluation.

[b] Coded as: -2 = explicitly negative, -1 = rather negative, 0 = mixed (both positive and negative), 1 = rather positive, 2 = explicitly positive.

[c] Coded as: -1 = negative (refers to “criticism” for AUTNES), 0 = neutral or mixed (both, positive and negative) (refers to “neutral” for AUTNES), 1 = positive (refers to “approval” in AUTNES).

[d] We ordered 10 judgments per sentence from CrowdFlower. We received 10 to 13 assessments per sentence, in sum 5,059 instead of 5,000. We included all to exhaust the crowd’s potential.

We asked the 5 offline coders to evaluate 500 sentences. Due to missing data (11 annotations) we received 2,489 evaluations instead of 2,500.

[e] The second item (evaluation tendency 5-point scale) was only displayed to those that selected ‘evaluation’ with regard to the first item (evaluation dichotomous).

[f] To compare crowdworkers, offline coders and AUTNES coders we selected only the ratings for the 196 sentences with concurring targets of opinion for this analysis.

p < .001. The results of the Games Howell post hoc test showed that the crowdworkers’ evaluations (M = -0.20, SD = 0.77, N = 1981) differed significantly from the offline coders’ judgments (M = -0.31, SD = 0.74, N = 979, p = .001) but not from those of the AUTNES coders, M = -0.14, SD = 0.69, N = 196, p = .481. We obtained almost identical results again when we replaced the unweighted CrowdFlower annotations with those weighted according to the trust scores, F(2, 546) = 8.10, p < .001.

In summary, these findings show that in spite of the homogenous coding instructions the judgments of the crowdworkers and those of the offline coders differ significantly, especially concerning the first part of the rating task (“Is there a rating of the target of opinion?”). In terms of the comparison with the AUTNES coding (obtained through trained expert coders with different instructions) we found that only the offline coders’ evaluation deviates significantly. Moreover, there were no significant differences between the unweighted and weighted crowdsourcing data.

Study 2: Experimental assessment of varying job designs

Data and methods

Overview

Rather than considering different data sources’ impact on reliability and validity (Study 1), Study 2 explores empirical concurrent validity of crowdworker data based on different job designs. For this experimental study, we manipulated three factors of the job design. More specifically, the way of material (here sentences) presentation, coding instructions and answer option formats used for the crowdsourcing and offline coding task in Study 1 were set as the “baseline” condition. Eleven additional experimental conditions varied then with regard to one of the factors holding the other two constant to be later contrasted with the “baseline”. They allow testing the impact of two alternative presentations of the material (referred to as M1 and M2), four different ways of giving coding instructions (I1–I4), and five alternative answer option formats (A1–A5). All 12 conditions are introduced in detail further below.

Procedure, crowdworkers, and quality control

The crowdworkers were randomly assigned to one of the 12 conditions and asked to evaluate 30 sentences. The sentences are a random sample of the 500 sentences rated for Study 1. For all conditions, the sentences were presented in random order and depicted together with the question(s) and scale(s) on separate pages.

In order to implement the randomization, we used SoSci Survey (https://www.soscisurvey.de) and CrowdFlower's survey option. Thus, we offered our job on CrowdFlower and redirected interested crowdworkers to SoSci Survey.

The CrowdFlower setting tools helped to again select only German speaking workers from Austria or Germany who are designated high quality workers. 588 crowdworkers participated between June 9 and June 30, 2016. Seventy-eight of them were excluded due to poor quality answers when responding to the test questions. It is not possible to use CrowdFlower's common quality control tools for jobs that direct the workers to external platforms. This is why we made the crowdworkers enter a job-specific code to receive payment once they had completed the job. We passed this code only to those that gave the correct answer to a test question[7] which was randomly mixed into the questionnaire. Workers that failed that test ($N = 59$) or did not take it because they canceled earlier ($N = 15$) were excluded. Extreme straight-liners ($N = 4$), workers that constantly selected no evaluation or the same combination of answers, were excluded. In summary, the assessments of 510 crowdworkers were used for the analysis.

Experimental conditions

Crowdworkers assigned to the "baseline" condition were confronted with sentences, instructions, and scales identical to those that had been used in Study 1. Hence the target of opinion was replaced by "Political Actor_X" or "Party_X", the coding instructions asked for explicit evaluations, perspective taking, and included the additional guiding comment "Please do not overuse neutral-no evaluation". Crowdworkers evaluated the sentences in a two-step process. For each sentence the first question asked about the general presence of an evaluation ("evaluation", "neutral-no evaluation"). If "evaluation" was selected a second question appeared and asked for the direction of the evaluation ("explicitly negative", "rather negative", "mixed [both, positive and negative]", "rather positive", "explicitly positive"). Generally, this design was kept constant in all other experimental conditions, except for the manipulation of one factor in each group. Table 4 presents the specific distinct features of each condition.

The *material presentation* was manipulated in two ways, that is regarding the specification of the target of opinion and anonymization of the target, to examine potential biased judgments due to workers' political preferences. Thus in contrast to the "baseline" condition where the target of opinion was fully replaced by "Political Actor_X" or "Party_X", the sentences in the two other experimental conditions used the original actor and party names, the first (M1) without, the second (M2) with special marking (X) of the target of opinion.

Regarding different *coding instructions* and contrasting the "baseline" instruction, we tested different question formats and the impact of a guiding comment. Crowdworkers in three additional conditions were either asked for '"explicit evaluations" (I1), simply for an "evaluation" (I2), or to "take the perspective of the target of opinion" (I3). The relevant part of the coding instruction was constantly repeated throughout the questionnaire and the manipulated words were displayed in bold. The instruction of another condition (I4), and again in contrast to the "baseline" did *not* include the additional guiding comment "Please do not overuse neutral-no evaluation".

Five additional experimental conditions considered various *answer option formats*. The "baseline" condition requested to rate the sentences in a two-step process. The two corresponding scales were flipped, the first scale in condition A1 ("neutral-no evaluation", "evaluation"), the second scale in condition A2 ("explicitly negative", "rather negative", "mixed [both, positive and negative]", "rather positive", "explicitly positive") and both scales in condition A3. In condition A4 the two scales were merged into one. The crowdworkers evaluated the sentences with the options "neutral-no evaluation", "explicitly positive", "rather positive", "mixed (both positive and negative)", "rather negative", and "explicitly negative". For A5 the options were further melted down to "positive", "neutral or mixed (both positive and negative)", and "negative". Again we stress that here the aim was to provide basic

[7] We pasted the remark "Important: This question is a test to check if the sentences are read carefully. Please select "I do not know", to pass the test. If you select "evaluation" or "neutral-no evaluation", you will have to finish the job without extra payment" directly after a sentence that was assumingly meant to be rated.

Table 4. Experimental manipulations of material presentation, coding instructions, and answer option formats [study 2].

Material presentation	Coding instructions		Answer option formats	
	Question formats	Guiding comment		
Baseline condition: Actor_X/Party_X	Look for an explicit evaluation; take the perspective of the evaluated subject;	Please do not overuse neutral-no evaluation	*Scale 1*: Evaluation Neutral-no evaluation	*Scale 2*: Explicitly positive Rather positive Mixed (both, positive and negative) Rather negative Explicitly negative
Other conditions: introduced in the column of the factor to be contrasted with the baseline, while the respective two other factors are held constant with the baseline				
M1: *Original* e.g. Werner Faymann /SPÖ	I1: Look for an explicit evaluation	I4: *Does not NOT include* 'Please do not overuse neutral-no evaluation'	A1: *Scale 1*: Neutral-no evaluation Evaluation	*Scale 2*: Explicitly positive Rather positive Mixed (both, positive and negative) Rather negative Explicitly negative
M2: *Original* (X) e.g. Werner Faymann (X)/SPÖ (X)	I2: Look for an evaluation		A2: *Scale 1*: Evaluation Neutral-no evaluation	*Scale 2*: Explicitly negative Rather negative Mixed (both, positive and negative) Rather positive Explicitly positive
	I3: Take the perspective of the evaluated subject		A3: *Scale 1*: Neutral-no evaluation Evaluation	*Scale 2*: Explicitly negative Rather negative Mixed (both, positive and negative) Rather positive Explicitly positive
			A4: Neutral-no evaluation Explicitly positive Rather positive Mixed (both, positive and negative) Rather negative Explicitly negative	
			A5: Positive Neutral or mixed (both positive and negative) Negative	

empirical evidence regarding the main effects of the different factors that were manipulated, and hence we did not consider the interplay of the factors.

Results

For the twelve experimental conditions, we collected 15,224[8] crowdcoded annotations. The number of cases per condition ranged from 1,020 (6.7%) for A3 to 1,440 (9.4%) for M1. Table 5 presents the sentences' mean ratings (and standard deviations) for all conditions.

Material presentation

The crowdworkers selected "evaluation" more frequently than "neutral-no evaluation" in the three conditions for with which we compared the manipulated anonymity of the target of opinion and target

[8] All 30 sentences were rated by 505 of 510 workers, hence 15,150 annotations. Five workers together annotated the remaining 74 sentences. Their judgements are part of the sample since they passed the test (see endnote 7) before they canceled the job.

Table 5. Ratings of crowdworkers (*N* = 510) for 12 conditions that vary with regard to material presentation, coding instructions and answer option formats [study 2].

	Evaluation dichotomous[a]		Evaluation tendency 5-point scale[b]		Evaluation tendency 3-point scale[c]	
Condition	Mean *(SD)*	*N*	Mean (*SD*)	*N*	Mean (*SD*)	*N*
Baseline	.60 (0.49)	1380	−0.37 (1.20)	821	−0.16 (0.68)	1380
M1	.67 (0.47)	1440	−0.38 (1.25)	970	−0.20 (0.72)	1440
M2	.65 (0.48)	1350	−0.40 (1.21)	873	−0.19 (0.71)	1350
I1	.62 (0.49)	1296	−0.30 (1.28)	797	−0.14 (0.71)	1296
I2	.63 (0.48)	1320	−0.39 (1.22)	834	−0.19 (0.69)	1320
I3	.66 (0.48)	1306	−0.29 (1.23)	857	−0.15 (0.71)	1306
I4	.58 (0.49)	1303	−0.36 (1.24)	755	−0.17 (0.69)	1303
A1[d]	.59 (0.49)	1230	−0.51 (1.25)	729	−0.22 (0.69)	1230
A2	.64 (0.48)	1049	−0.36 (1.21)	672	−0.17 (0.69)	1049
A3	.67 (0.47)	1020	−0.49 (1.26)	688	−0.22 (0.73)	1020
A4[d,e]	.76 (0.42)	1200	−0.32 (1.26)	917	−0.19 (0.77)	1200
A5					−0.11 (0.82)	1330

[a] Coded as: 1 = evaluation, 0 = neutral-no evaluation.
[b] Coded as: -2 = explicitly negative, -1 = rather negative, 0 = mixed (both positive and negative), 1 = rather positive, 2 = explicitly positive.
[c] Coded as: -1 = negative, 0 = neutral or mixed (both, positive and negative), 1 = positive.
[d] Recoded to a 3-point scale for a comparison with A5.
[e] Recoded to two scales for a comparison with A1.

specification. An evaluation was perceived by 67% of the crowdworkers when the target of opinion was named without any specific labeling (M1), by 65% when the actor or party was marked with (X) (M2), and finally by 60% when the target of opinion was kept anonymous ("baseline"). These differences point to a significant, albeit weak relationship between the workers' assessment if a political actor was evaluated or not and the material presentation, $\chi 2(2, N = 4170) = 19.44$, $p < .001$.

Regarding the evaluative tendencies (5-point scale ranging from -2 = explicitly negative to 2 = explicitly positive) the workers' assessment in the 'baseline' condition, in M1 and M2 were overall slightly negative, "baseline": *M* = -0.37, *SD* = 1.20; M1: *M* = -0.38, *SD* = 1.25; M2: *M* = -0.40, *SD* = 1.21. The marginal differences between the conditions are not significant, *F*(2, 2661) = 0.16, *p* = .851. Thus, the assessment if a perceived evaluation is positive or negative appears to be independent from the sentences' manipulated degree of anonymity and target specification.

Coding instructions

Sixty-two percent of the workers selected "evaluation" when they were repeatedly and exclusively asked to look for "explicit evaluations" (I1), 66% when "perspective taking of the evaluated person" was constantly and again exclusively stressed (I3), 60% when both ("baseline") and 63% when just "evaluations" (I2) was mentioned. Overall, question wording appeared to matter; we found a significant but weak relationship between the decision to select "evaluation" or "neutral-no evaluation" and the question format, $\chi 2(3, N = 5302) = 11.54$, *p* = .009. With regard to the tendency of evaluation, differences between the four conditions ("baseline", I1, I2, I3) were only marginal, *F*(3, 3305) = 1.30, *p* = .274. The exclusion of the guiding comment "Please do not overuse neutral-no evaluation" (I4) did not lead to significantly different ratings which regard to the question if ($\chi 2[1, N = 2683] = 0.66$, *p* = .415) and how (*t*[1574] = -0.06, *p* = .952) the target of opinion is evaluated. The differences between the crowdworkers' annotations in the "baseline" condition, where the request was included, and in condition I4, where it was missing, are marginal and not significant.

Answer option formats

When "evaluation" was the upper and "neutral-no evaluation" the lower option to select ('baseline' and A2, *N* = 2429), 62% of the crowdworkers chose "evaluation", compared to 63% of the crowdworkers in the conditions A1 and A3 (*N* = 2250), where the two answer options were flipped. This difference is not significant, $\chi 2(1, N = 4679) = 1.14$, *p* = .287. In terms of the evaluation tendency, the annotations did not differ significantly ($t[1491] = -0.15$, $p = .880$) between those workers who got "explicitly positive" as first option (baseline) ($M = -0.37$, $SD = 1.20$) and those who got "explicitly negative" in the top position (A2) ($M = -0.36$, $SD = 1.21$); both perceive on average a slightly negative evaluation. In the conditions where the first scale was already flipped with "neutral-no evaluation" on top for both (A1 and A3), the same finding was made for their inversely flipped evaluation tendency scales ($t[1415] = -0.37$, $p = .712$); the evaluation tendency was not significantly different in A1 ($M = -0.51$, $SD = 1.25$) and A3, $M = -0.49$, $SD = 1.26$.

Other scale variations were tested with A4 and A5, where the crowdworkers rated the sentences with just one scale instead of two. Condition A4 included identical answer options as the 'baseline' but combined into one scale and therefore had six answer options. In order to compare the annotations of A4 with A1, the most similar condition ("neutral-no evaluation" is the first presented option), the answers for A4 were recoded. "Evaluation" was selected by 59% when it is one out of two options (A1) and indirectly by 76% when five options out of six refer to an evaluation (A4), a statistically significant difference ($\chi2[1, N = 2430] = 81.73$, $p < .001$). Furthermore, the evaluation of the political actor or party was perceived significantly less negative in A4 ($M = -0.32$, $SD = 1.26$) compared to workers in A1 ($M = -0.51$, $SD = 1.25$), $t(1644) = -3.14$, $p = .002$. A possible explanation is that the additional "neutral-no evaluation" option for A4 shifted the balance of the original 5-point scale.

We compared the ratings of A5, given on a 3-point scale with "positive", "neutral or mixed (both, positive and negative)" and "negative", with those of A1 and A4. Both were recoded accordingly. Again, the answer option formats seemed to significantly influence the workers' rating, $F(2, 2493) = 6.40$, $p = .002$. Negative ratings were most frequent ($M = -0.22$, $SD = 0.69$) when "negative" was referred to in two responses out of five options on the second scale (A1), followed by the condition ($M = -0.19$, $SD = 0.77$) where a negative evaluation was represented by two out of six options on a single scale (A3), and finally the condition ($M = -0.11$, $SD = 0.82$) where "negative" was one out of three options on a single scale (A4).

In summary, workers appear to be overall rather robust against minor changes with regard to the coding instructions. Only when it comes to larger scale variations, and somewhat less pronounced regarding different material presentations, the workers rate the same sentences differently. We proceed to discuss these results in more detail and consider implementations for the use of crowdsourcing for this type of content analysis task and beyond.

Overall conclusion

The aim of this article was twofold. On the one hand, we sought to systematically compare the reliability and validity of content analytical data generated by a crowdsourcing platform with that coming from more controlled, traditional and offline settings. On the other hand, we were interested whether differences in task presentation within a crowdsourcing environment, ranging from material and instructions to answer scales, would yield significantly different results. We focused on the coding of evaluations of political actors in sentences from news articles as a critical case, given the rather subjective and less manifest nature of such assessments. To address these issues, two separate empirical studies were conducted.

Regarding reliability and empirical validity of crowdcoded data (Study 1), the results suggest that the crowdworkers did not perform particularly poorly compared to offline coders in a controlled setting who relied on identical coding instructions. Values for ICR did not differ substantially between the crowdworkers and the offline coders. It is noteworthy, however, that reliability scores do improve (if only slightly) when only the most trusted crowdworkers are selected. This points to a potential usefulness of relying on data aggregation based on trust scores when seeking more reliable data. Given the advantages of crowdcoded data in terms of being "intrinsically reproducible" (Benoit et al., 2016, p. 278) at a higher speed and lower costs, smaller deviations in ICR may be a reasonable price to pay. In addition it needs to be taken into account that crowdcoding procedures would produce data from several coders on *all* cases. Assuming validity of the data aggregations (as discussed below) a stronger variation around the aggregated value, and thus a lower ICR score, may not be as worrisome here as in a traditional setting.

Having said so, we certainly do acknowledge that reliability of the first step, identifying the presence of an evaluation, is very low and according to common standards not acceptable in both data sources. It should be noted that we put the bar quite high for our coders, since "objectivity is a much tougher criterion to achieve with latent than with manifest variables, and for this reason, we expect variables measuring latent content to generally receive lower reliability scores." (Neuendorf, 2002, p. 146). Considering this, we are in fact quite pleased with the reliability scores on the second question, the evaluative tendency although also these scores fall well below common standards for high reliability. Equally important with regard to crowdsourcing data is the notion that crowds generally appear to perform less well when it comes to nominal multiple choice (such as yes-no) decisions (e.g. Prelec, Seung, & McCoy, 2017). This is well reflected in the differences between unacceptable and, albeit low but more acceptable reliability scores in the two different coding steps. Clearly, instructions to identify the presence of an evaluation need refinement, which is also reflected in the validation procedure. Future studies may address likelihood ratings regarding the presence of an evaluation as alternative.

Concerning empirical validity (Krippendorff, 2013), the second quality criterion we use to assess the crowdcoded data, we find noteworthy differences when comparing the outcomes of the crowdcoded data with those of our offline coders, in particular in terms of whether an evaluation was marked as present or not, but also with regard to the type of evaluation. The problem in qualitatively addressing this difference lies in the lack of a gold standard. Approximating such gold standard would be the AUTNES data (Eberl et al., 2016) which showed high reliability scores and has been extensively used for publications (e.g. Eberl, Boomgaarden, & Wagner, 2015; Eberl, Wagner, & Boomgaarden, 2017; Song, Nyhuis, & Boomgaarden, forthcoming). When contrasting the evaluations of the five offline coders and those of the crowdworkers to the AUTNES data, only the offline coders are shown to deviate significantly from the other groups. Thus, we see more overlap between the wisdom of the crowd and a set of experienced and trained coders using extensive coding instructions. This highly speaks for the appropriateness of using crowdsourcing as a tool for such content analysis. Given that presented test questions in a crowdsourcing job can be perceived of as at least some kind of coder training, one explanation for these findings is offered by the fact that offline coders were the only group that did not receive any training at all.

The findings presented in Study 1 are based on a sample of up to 500 sentences mentioning political actors, drawn from election and non-election periods and a great variety of newspapers. Given the breadth of this sample, we have little reason to doubt that the results would be generalizable to evaluations of actors in journalistic, and potentially other sorts of texts more generally. Crowd wisdom is used to assess the greatest variety of matters (e.g., Saur-Amaral, 2012) and if it works for judgments of evaluative tendency here, we do not see why it should not be used for the assessment of content characteristics of other kinds, ranging from certainly manifest to more projective latent (Potter & Levine-Donnerstein, 1999), subjective (Haselmayer & Jenny, 2014), or interpretative constructs. Despite this somewhat encouraging picture, it is important to consider whether variations in how material and coding instructions are presented would matter.

Our experimental findings (Study 2) overall suggest that minor changes in the presentation of a crowdsourcing task do not significantly influence the actual ratings of the crowdworkers. This is particularly true for the actual evaluative tendency of the coding. More specifically, we showed that different ways of asking for a sentiment annotation (e.g., "explicit evaluation" or "perspective taking") do not considerably affect the results. The same holds true for the request "Please do not overuse neutral-no evaluation". Moreover, our results suggest, that it is not worth the effort to anonymize the target of opinion. The minor influence of whether or not the political actor was named or made anonymous suggests either that the workers approach the evaluation task rather from a neutral perspective, that the sample overall is rather diverse in its political preferences, or that knowledge of politics was too low to affect the results. Furthermore, while larger scale modifications (one scale vs. two scales; the directionality and number of answer options) have an effect on the workers' ratings, simply flipping scales does not.

By contrast we saw a few more instances in which material presentation and instructions affected whether an evaluation was considered to be present or not. Anonymizing the material led to fewer identified evaluations, while the inclusion of "perspective taking" and the shift toward a combined measurement led to more identified evaluations. Researchers need to be sensitive to these issues. It

Table 6. Initial best practice recommendations for the use of crowdsourcing platforms for quantitative content analysis [study 1 and 2].

Validity and Reliability
Assess data reliability for all cases. If reliability scores are low, assess empirical validity by comparing aggregated data to subsample, manual coded gold standard.
Data aggregation weighted by workers' trust scores should be preferred above non-weighted aggregation.
Crowdcoded content data may produce more reliable and valid results for scale ratings than for nominal answer options.
Do not anonymize targets of opinion, variation in your crowd will prevent non-valid responses.
Keep tasks separate; try not to save money by demanding different judgments in one step.
Quality control
Apply test questions to monitor crowdworkers` performance on your job and to be able to sort out those who work poorly.
Apply test questions that are representative for your task to train your crowdworkers while they are working on the job.
Check if there are workers who constantly select the same combination of answers (straightliners).

Note. For many more questions that might come up when working with CrowdFlower we can recommend the CrowdFlower guides, documentations, and customer service.

appears that in particular the practice of reducing the task to one question only, in which the presence and evaluative tendency are measured at once, leads to considerably different outcomes. While this may be beneficial from a resource point of view, it provides for less valid findings. Overall we see that the first step, assessing whether an evaluation is present at all or not, appears to be more challenging than the actual coding of the evaluative tendency. Future research needs to take this into account and consider more appropriate instructions. Regarding our experimental findings we again have little reason to doubt their generalizability to other types of actor evaluations. Language may play a role here, and given our data were generated by German language instructions and German language material, it may be advisable to look at instructions and materials in other languages, preferably in a comparative manner.

In sum we believe our conclusions favorably speak to the potential of using crowdcoding as an attractive alternative for the costly and labor intensive traditional manual approach. Researcher, however, as with traditional content analysis, need to be sensitive to issues of reliability and validity, and give the novelty of the approach a lot needs to be done to work towards more standardization. Based on the first steps presented here, Table 6 provides some best practice recommendations.

While automated procedures increasingly replace human coders in content analysis procedures, real people are still needed to supplement or feed the machines, which is especially the case when latent content is to be coded. Here the advantages of crowdsourcing tools should lead to a serious consideration of these as an alternative to employing and training human coders. Such data, even with less than perfect reliability scores (e.g., Burscher, Odijk, Vliegenthart, De Rijke, & De Vreese, 2014), can be important tools for algorithm development in machine learning procedures. Furthermore, crowdsourcing offers the opportunity to reveal group decisions that may depict more valid results than the judgment of one single trained expert coder and also may be easier to replicate. Bearing in mind that in communication science we are often interested in the possible effects of news content on a broader public and not on trained experts, the judgments of lay crowdworkers may be more appropriate to use. This contribution makes first, but in our view important, steps towards a more reflective use of crowdcoding for content analysis purposes and hopefully sparks some sort of standardization of its procedures.

Funding

This research is conducted under the auspices of the Austrian National Election Study (AUTNES), sponsored by the Austrian Science Fund (FWF): S10908-G11.

References

Banducci, S., de Vreese, C., Semetko, H., Boomgarden, H., & Luhiste, M. (2010). *EES Longitudinal Media Study Data Advance Release Documentation*, version 15.10. 2010. Retrieved from www.piredeu.eu

Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods*, *43*(3), 800–813. doi:10.3758/s13428-011-0081-0

Benoit, K., Conway, D., Lauderdale, B. E., Laver, M., & Mikhaylov, S. (2016). Crowd-sourced text analysis: Reproducible and agile production of political data. *American Political Science Review*, *110*(2), 278–295. doi:10.1017/S0003055416000058

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis*, *20*(3), 351–368. doi:10.1093/pan/mpr057

Boumans, J. W., & Trilling, D. (2016). Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. *Digital Journalism*, *4*(1), 8–23. doi:10.1080/21670811.2015.1096598

Bradburn, N. M., & Sudman, S. (1979). *Improving interview method and questionnaire design – Response effects to threatening questions in survey research*. San Francisco, CA: Jossey-Bass Publishers.

Budak, C., Goel, S., & Rao, J. M. (2016). Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, *80*(S1), 250–271. doi:10.1093/poq/nfw007

Burscher, B. (2016). *Machine learning-based content analysis automating the analysis of frames and agendas in political communication research* (Doctoral dissertation). Faculty of Social and Behavioral Sciences, University of Amsterdam, Amsterdam, The Netherlands.

Burscher, B., Odijk, D., Vliegenthart, R., De Rijke, M., & De Vreese, C. H. (2014). Teaching the computer to code frames in news: Comparing two supervised machine learning approaches to frame analysis. *Communication Methods and Measures*, *8*(3), 190–206. doi:10.1080/19312458.2014.937527

Catallo, I. (2015). *Achieving quality in crowdsourcing through task design and assignment* (Doctoral dissertation). Dipartimento di Elettronica, Informazione e Bioingegneria, Polytechnic University of Milan, Milan, Italy.

Eberl, J.-M., Boomgaarden, H. G., & Wagner, M. (2015). One bias fits all? Three types of media bias and their effects on party preferences. *Communication Research*. doi:10.1177/0093650215614364

Eberl, J. M., Vonbun, R., Haselmayer, M., Jacobi, C., Kleinen-von Königslöw, K., Schönbach, K., & Boomgaarden, H. G. (2016). AUTNES manual content analysis of the media overage 2013. *GESIS Data Archive, Cologne*. [ZA5864 Data File Version 1.0.0]. doi:10.4232/1.12565

Eberl, J. M., Wagner, M., & Boomgaarden, H. G. (2017). Are perceptions of candidate traits shaped by the media? The effects of three types of media bias. *The International Journal of Press/Politics*, *22*(1), 111–132. doi:10.1177/1940161216674651

Finnerty, A., Kucherbaev, P., Tranquillini, S., & Convertino, G. (2013, September). Keep it simple: Reward and task design in crowdsourcing. In *Proceedings of the biannual conference of the Italian chapter of SIGCHI* (pp. 14:1–14:4). ACM. doi:10.1145/2499149.2499168

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297. doi:10.1093/pan/mps028

Hasegawa-Johnson, M., Cole, J., Jyothi, P., & Varshney, L. R. (2015). Models of dataset size, question design, and cross-language speech perception for speech crowdsourcing applications. *Laboratory Phonology*, *6*(3–4), 381–431. doi:10.1515/lp-2015-0012

Haselmayer, M., & Jenny, M. (2014). *Measuring the tonality of negative campaigning: Combining a dictionary approach with crowd-coding*. Paper presented at political context Matters: Content analysis in the social sciences, Mannheim, Germany.

Haselmayer, M., & Jenny, M. (2016). Sentiment analysis of political communication: Combining a dictionary approach with crowdcoding. *Quality & Quantity*, 1–24. doi:10.1007/s11135-016-0412-4

Heise, D. R. (2010). *Surveying cultures: Discovering shared conceptions and sentiments*. Hoboken, NJ: John Wiley & Sons.

Hill, C. A., Dean, E., & Murphy, J. (2014). *Social media, sociality, and survey research*. Hoboken, NJ: John Wiley & Sons.

Howe, J. (2006). The rise of crowdsourcing. *Wired Magazine*, *14*(6), 1–4.

Hsueh, P. Y., Melville, P., & Sindhwani, V. (2009, June). Data quality from crowdsourcing: A study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing* (pp. 27–35). Association for Computational Linguistics, Morristown, NJ, USA.

Jacobi, C., van Atteveldt, W., & Welbers, K. (2016). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, *4*(1), 89–106. doi:10.1080/21670811.2015.1093271

Kazai, G., Kamps, J., Koolen, M., & Milic-Frayling, N. (2011, July). Crowdsourcing for book search evaluation: Impact of hit design on comparative system ranking. In *Proceedings of the 34th international ACM SIGIR conference on research and development in information retrieval* (pp. 205–214). ACM. doi:10.1145/2009916.2009947

Kittur, A., Chi, E. H., & Suh, B. (2008, April). Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 453–456). ACM. doi:10.1145/1357054.1357127

Kittur, A., Nickerson, J. V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., … Horton, J. (2013, February). The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work* (pp. 1301–1318). ACM. doi:10.1145/2441776.2441923

Kleinen-von Königslöw, K., Eberl, J. M., Haselmayer, M., Jacobi, C., Vonbun, R., Boomgarden, H. G., & Schönbach, K. (2016). *AUTNES manual content analysis of the media coverage 2013 documentation version 1.0.0*. Cologne, Germany: GESIS Data Archive.

Komarov, S., Reinecke, K., & Gajos, K. Z. (2013, April). Crowdsourcing performance evaluations of user interfaces. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 207–216). ACM. doi:10.1145/2470654.2470684

Krippendorff, K. (2013). *Content analysis. An introduction to its methodology*. Thousand Oaks, CA: Sage.

Levay, K. E., Freese, J., & Druckman, J. N. (2016). The demographic and political composition of mechanical turk samples. *SAGE Open*, *6*, 1–17. doi:10.1177/2158244016636433

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, *5*(1), 1–167. doi:10.2200/S00416ED1V01Y201204HLT016

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research*, *28*(4), 587–604. doi:10.1111/j.1468-2958.2002.tb00826.x

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's mechanical turk. *Behavior Research Methods*, *44*(1), 1–23. doi:10.3758/s13428-011-0124-6

Mohammad, S. M. (2016a). Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In H. L. Meiselman (Ed.), *Emotion measurement* (pp. 201–238). Duxford/Kidlington, UK: Elsevier Ltd.

Mohammad, S. M. (2016b). A practical guide to sentiment annotation: Challenges and solutions. In *Proceedings of the workshop on computational approaches to subjectivity, sentiment and social media analysis*. Retrieved from http://www.saifmohammad.com/WebDocs/SentimentAnnotation-wassa2016.pdf

Mullinix, K. J., Leeper, T. J., Druckman, J. N., & Freese, J. (2015). The generalizability of survey experiments. *Journal of Experimental Political Science*, *2*(2), 109–138. doi:10.1017/XPS.2015.19

Neuendorf, K. A. (2002). *The content analysis guidebook*. Thousand Oaks, CA: Sage.

Oosterman, J., Yang, J., Bozzon, A., Aroyo, L., & Houben, G. J. (2015). On the impact of knowledge extraction and aggregation on crowdsourced annotation of visual artworks. *Computer Networks*, *90*, 133–149. doi:10.1016/j.comnet.2015.07.008

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2* (1), 1–90. doi:10.1561/1500000011

Peer, E., Samat, S., Brandimarte, L., & Acquisti, A. (2016). Beyond the turk: An empirical comparison of alternative platforms for crowdsourcing online behavioral research. *Social Science Research Network*. doi:10.2139/ssrn.2594183

Potter, W. J., & Levine-Donnerstein, D. (1999). Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research*, *27*(3), 258–284. doi:10.1080/00909889909365539

Prelec, D., Seung, H. S., & McCoy, J. (2017). A solution to the single-question crowd wisdom problem. *Nature*, *541* (7638), 532–535. doi:10.1038/nature21054

Riloff, E., & Wiebe, J. (2003). Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on empirical methods in natural language processing* (pp. 105–112). EMNLP. doi:10.3115/1119355.1119369

Saur-Amaral, I. (2012, January). Wisdom-of-the-crowds to enhance innovation: A conceptual framework. In *ISPIM conference proceedings* (pp. 1–7). ISPIM, Barcelona, Spain.

Saxton, G. D., Oh, O., & Kishore, R. (2013). Rules of crowdsourcing: Models, issues, and systems of control. *Information Systems Management*, *30*(1), 2–20. doi:10.1080/10580530.2013.739883

Schönbach, K., Kleinen-von Königslöw, K., Eberl, J. M., Haselmayer, M., Jacobi, C., & Vonbun, R. (2016). *AUTNES manual content Analysis of the media coverage 2013 – codebuch*. Cologne, Germany: GESIS Data Archive.

Schwarz, N., Knäuper, B., Oyserman, D., & Stich, C. (2008). The psychology of asking questions. In E. D. De Leeuw, & D. A. Dillman (Eds.), *International handbook of survey methodology* (pp. 18–22). New York, London: Taylor and Francis Group.

Shank, D. B. (2016). Using crowdsourcing websites for sociological research: The case of amazon mechanical Turk. *The American Sociologist*, *47*(1), 47–55. doi:10.1007/s12108-015-9266-9

Simpson, E. D., Venanzi, M., Reece, S., Kohli, P., Guiver, J., Roberts, S. J., & Jennings, N. R. (2015, May). Language understanding in the wild: Combining crowdsourcing and machine learning. In *Proceedings of the 24th international conference on world wide web* (pp. 992–1002). ACM, New York, NY.

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008, October). Cheap and fast—But is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 254–263). Stroudsburg, PA: Association for Computational Linguistics.

Song, H., Nyhuis, D., & Boomgaarden, H. G. (forthcoming). A network model of negative campaigning: The structure and determinants of negative campaigning in multi-party systems. *Communication Research*.

Surowiecki, J. (2005). *The wisdom of crowds*. New York, NY: Anchor.

Vakharia, D., & Lease, M. (2015, March). Beyond mechanical turk: An analysis of paid crowd work platforms. In *Proceedings of the iConference*. Newport Beach, CA: iSchools organization and the Donald Bren School of Information and Computer Sciences at the University of California, Irvine. Retrieved from https://www.ideals.illinois.edu/bitstream/handle/2142/73639/138_ready.pdf?sequence=2&isAllowed=y

Vuurens, J., de Vries, A. P., & Eickhoff, C. (2011, July). How much spam can you take? an analysis of crowdsourcing results to increase accuracy. In *Proc. ACM SIGIR workshop on Crowdsourcing for Information Retrieval (CIR'11)*. Retrieved from http://mediamatica.ewi.tudelft.nl/sites/default/files/paper_2.pdf

Wang, S., Huang, C. R., Yao, Y., & Chan, A. (2015, October). Mechanical Turk-based experiment vs laboratory-based experiment: A case study on the comparison of semantic transparency rating data. In *29th Pacific Asia conference on language, information and computation* (pp. 53–62). Shanghai, China: Department of Computer Science and Engineering, Shanghai Jiao Tong University.

Weiß, H.-J., Maurer, T., Schwotzer, B., Trautmann, H., & Zhu, J. (2009). *Fernsehnachrichtenanalyse zum Bundestagswahlkampf 2005. Methodenbericht/Methodendokumentation*. Potsdam, Germany: GöfaK Medienforschung GmbH.

Zhu, D., & Carterette, B. (2010, July). An analysis of assessor behavior in crowdsourced preference judgments. In *SIGIR 2010 workshop on crowdsourcing for search evaluation* (pp. 17–20). New York, NY: ACM.