Can the online-crowd match real expert judgments? How task complexity and coder location affect the validity of crowd-coded data (new version of a paper/research note prepared for the European Political Science Association's 2017 conference in Milano, Italy)

By: Dr Alexander Horn, Assistant Prof., Aarhus University, Denmark, M: ahorn@ps.au.dk

Abstract

Crowdcoding is a novel technique that allows for fast, affordable, and reproducible online categorization of large numbers of statements. It combines judgements by multiple, paid, non-expert coders to avoid miscoding(s). Benoit et al. (2016) argue that crowdcoding could replace expert judgements; using the coding of political texts as an example in which both strategies produce similar results. Since crowdcoding yields the potential to extend the replication standard to data production and to "scale" coding schemes based on a modest number of carefully devised test questions and answers, it is important that we better understand its possibilities and limitations. While previous results for low complexity coding tasks are encouraging, we assess whether and under what conditions simple *and* complex coding tasks can be outsourced to the crowd without sacrificing content validity in return for scalability. The simple task is to decide whether a party statement counts as positive reference to a concept – in our case equality. The complex task is to distinguish between five concepts of equality. To account for the crowdcoder's contextual knowledge, we vary the IP restrictions. The basis for our comparisons are 1404 party statements; coded by experts and the crowd (resulting in 30.000 online judgements). We compare the expert-crowd match at the statement- and party level and find that the (aggregated) results are substantively similar even for the complex task, suggesting that complex category schemes can be scaled via crowdcoding. The match is only slightly higher when IP restrictions are used as an approximation of coder expertise.

1. Introduction

In this paper we assess how the results of online crowdcoding compare with the results of codings by experts and how the match between both varies depending on the complexity of the coding task and coder characteristics such as the geographical location. Our specific aim is to identify conditions under which crowdcoding can replace expert judgements. The broader aim is to discuss the limits and possibilities of this fascinating new part of the tool box of political scientists. Thus, the paper targets a broad audience of empirical researchers rather than methodologists only. Crowdcoding is a novel data gathering technique that allows for fast, affordable, reproducible and – at least potentially – very fine-grained online categorization of a very large number of items. It combines multiple judgements (per codeable unit) by paid non-expert coders and their trust scores to avoid miscoding(s). Coders must first pass an entry quiz to qualify and are continuously screened with further test questions to deselect underperformers throughout the actual coding process. Crowdcoding can be regarded as a specific form of crowd sourcing and has its roots in the idea that multiple independent judgements by a diverse crowd of contributors (in our case coders) can – when aggregated (sic!) – match or outperform experts and their judgements. This simple but powerful idea that good collective decisions can emanate from various averaged independent judgements of non-experts is long discussed in academia, business, and popular science (see Surowiecki, 2004; Lehman and Zobel, 2017). Yet, notwithstanding instructive earlier studies with positive conclusions regarding the validity of crowdcoded data (e.g., Berinsky et al., 2014; Haselmayer and Jenny, 2016), it seems fair to say that crowdcoding is only starting to gain traction in political science at large since Benoit et al. (2016) have convincingly argued that the results of expert judgements – still considered the gold standard by many (for instance when it comes to the location of parties) – can be matched with crowdcoding; at least for simple coding tasks. This is

significant, since experts are expensive and in short supply and automated (coding) methods are not yet good enough at extracting meaning (ibid, 280).

One important limitation of this and other previous studies – and our point of departure – is that complex coding tasks and the conditions under which they yield valid codings are not considered. In our view, the reason why the nascent debate about crowdcoding in political science should not just concern methodologists, but is of interest to most political scientists, is not the prospect of reproducibility and scalability *per se*, but the possibility to move beyond the "off-the-shelf" data sets that limit the scope of topics researched and the validity of the existing proxy measures. Previously, researchers who had in mind research questions and theoretical concepts for which the indicators in canonical large-N datasets were inappropriate (often, but not always, this would mean too broad) had three options: dropping the specific research question, starting an expensive long-term data-infrastructure project, or settling for suboptimal – often too unspecific – proxies to operationalize the theoretical concepts of interest. Crowdcoding could help them to scale up a new category scheme that better reflects the concepts they are interested in, without sacrificing reliability. So, ideally, the trade-offs that empirical researchers face between reliability and content validity *could* be mitigated using crowdcoding. Importantly, this does not mean that we should dismiss canonical data-sets (they are indispensable to cumulate evidence), but rather to complement and refine them.

One example may help to illustrate this point. Although it is widely acknowledged that inequality is one of the – if not *the* – political challenge of the 21st century and despite renewed interest in the politics on inequality (Jensen and van Kersbergen, 2016), we do not have large-N data on parties and governments' conceptions of and positions on equality and inequality. It makes a crucial difference whether parties prioritize equality of outcome, equality of opportunity, or something else (Horn et al. 2017). Yet, the canonical data based on party manifestos "only" tells us how often

equality is referred to (positively). If we are interested in Amartya Sen's classic *Equality of What*-question (1979) and the effect that different concepts of (and positions on) equality have on political outputs (policy) and outcomes (inequalities), we have to better understand *what* parties talk about *when* they talk about equality and inequality. Whether such a specification of broad and simple measures into more exact and theoretically meaningful subcategories can be achieved via crowdcoding is an open question that we will address in the remainder of this paper. Since the coded manifestos are now available as digitized text corpus at the level of (quasi-) sentences (Merz et al., 2016), we can use crowdcoding to take stock of the underlying variation. Since this variation has already been mapped for selected German manifestos (Horn et al., 2017), this gives us the opportunity to compare the results obtained from these expert judgements with the results from crowdcoding across six different scenarios. We vary two factors: The first is task complexity. It is low when coders simply have to decide whether a party statement is positively related to equality and high when they also have to decide which out of 5 concepts of equality is addressed in the statement. The second factor can be called *coder expertise* or *contextual knowledge*. A priori, it is conceivable that knowledge of the subject(s) (here German parties) makes coding easier, but also that it introduces biases. Coder expertise is lower when there are no geographical criteria for the selection of coders and higher when there is a strong restriction (a German IP address). Since this initial setup conflates German and non-German IPs, we have added a third setting in which only coders with a *non*-German IP are allowed to participate.[1]

Combining the two complexity levels and the three IP settings results in six coding scenarios. Strictly speaking, we should at best speak of *higher* expertise or rather more contextual knowledge, as the online coders are clearly not experts. We think that (real) experts are political scientists,

[1] The language criteria in CrowdFlower can be ignored by the coders, though of course only those with a fairly good command of German will pass the qualification test. We want to know whether context matters; and thus use the IP.

familiar not just with the country, but also with the policy field(s) under study[2]. The comparison with the expert codings and the resulting *match* is crucial, as we would otherwise have no yardstick to assess the content validity (see Gerring, 2012) of the crowd coded results.

We will first explain the simple and the complex category scheme and how we calculate the match between the results of the expert judgements and the results of crowdcoding. Then we explain the practical implementation of our six different scenarios in the leading crowdcoding platform CrowdFlower.com. Afterwards we will present the results at the statement and the party-level. Finally, we draw conclusions regarding the possibilities and limits of crowdcoding in political science; especially when the researcher works with a complex category scheme. We find that crowdcoding can yield results similar to expert codings – even for more complex tasks. This is true independent of IP restrictions. The expert-crowd match is only modestly higher for German IPs.

2. Assessing the expert-crowd match at different levels of task complexity and coder expertise

To assess in as detailed a manner as possible whether experts can be replaced by crowd coders we need a fully coded corpus of expert judgements. In principle – because aggregation supposedly renders differences at the level of individual codings irrelevant – it may be sufficient to compare the result. In our case, this would be the shares of different subcategories of equality – or simply the dominant type of equality that parties talk about. Miscodings of one coder do not necessarily matter (we call this level of individual coding decisions *coder level*) when four others code the same statement correctly. It is also true that even statements miscoded after all coders have given their judgement on one item (we are now at the *statement leve*l) may not matter, as these mistakes could

[2] During the WZB's Manifesto User conference 2015 in Berlin a participant reported that the first thing he did when asked about expert judgements on party positions in the country where he is located was to look up the Manifesto data. This anecdote illustrates that even an accomplished political scientist familiar with the general country context may not necessarily be familiar with the details of the debates regarding migration, welfare, law- and order, or economic policy.

even out later when we compare conceptions of equality across parties (the *party level*). Theoretically, it is conceivable that a high expert-crowd match of the party priorities could emerge without a high match between experts and the crowd at the level of coded statements. However, we think in order to put trust in the method of crowdcoding, one must be able to trace why and how its results match or do not match the results obtained from experts. In other words, while it is the core idea of crowdcoding that single miscodings at the coder level do not matter after aggregation, we would still expect a basic match at the statement level, even though it is the party level that ultimately matters.

We use as a yardstick the codings for the 1404 party statements (including 66 test questions) on equality coded by Horn et al. (2017). These 1404 statements are from the party programs of the last German general election (in 2013) and represent all statements devoted to equality (according to the coders of the Manifesto project) by the five parties that were in the Bundestag pre-election. The experts discussed controversial categorizations to reach agreement, introducing a subjective element. This very subjectivity is inherent to expert judgements and in our view part of the motivation behind and appeal of combining oft-repeated codings by non-experts (e.g., in Benoit et al., 2016). So while we need the codings of the experts as a yardstick to assess the expert-crowd match, we by no means think that experts are infallible. This reservation notwithstanding, said analysis shows that there is a considerable left-right gradient in what parties talk about when they talk about equality. Left parties emphasize economic equality and (centre-)right parties equal chances and opportunities. The Gretchen question is whether we – or rather the crowd – can reproduce the results of this coding exercise by experts despite the complex category scheme? Table 1 and 2 entail the coding scheme in its simple and its complex form. Of course, the instructions for the coders were in German – as were the party statements (see Appendix 1 for the German version).

Table 1: Simple task: template used to brief coders and the category scheme they see

Does the following statement from a party program exert a positive reference to equality, and/or social justice, and/or equal treatment of all people? If YES, please select category 1 /yes? If NO, please select category 2 /no. 1) **Yes** 2) **No**

Table 2: Complex task: template used to brief coders and the category scheme they see

Does the following statement from a party program exert a positive reference to equality, social justice, and/or equal treatment of all people? If YES, which of the categories 1 to 5 fits best? I NO, please select category 6. Please read all of the 6 category descriptions carefully! If you are of the opinion that more than one category is relevant, please decide which category is emphasized most clearly. 1) **Economic equality** (for example: we criticize that the little guy is doing badly; we criticize that economic inequality in our country is on the rise; we must redistribute more from top to bottom; strong shoulders must carry more; society is drifting apart; the gap between rich and poor must narrow again; wealth tax now; we have to distribute the gains from globalization in a just manner; property for everyone equals social peace). 2) **Mentioning** of equality, (social) justice, and solidarity – but without getting concrete (for example: we are the party of justice; we stand for (more) solidarity. 3) **Equality of chances and social mobility:** (for example: the education system must be more permissive; more children of workers must make it to university; social background must not decide over the fate/future of children; everyone must have a chance – independent of the parents' purse). 4) **Inclusion, nondiscrimination, antidiscrimination**: (for example: more must be done against the discrimination of woman, homosexuals, foreigners, disabled people, the old; our party stands for diversity and inclusion; we want a colorful society; no one must be discriminated). Statements on fair wages for woman and against the gender pay gap also belong into this category 4. 5) **Other:** there is a link to equality, (social) justice and equal treatment of all people, but the statements fits into none of the previous four categories (for instance: global justice; development aid; internet justice; equal access to the internet for all; Aufbau Ost/support of the new Laender; mobility for all; energy justice; climate- und environmental justice, housing related justice). 6) **No – not a positive reference**. Examples for irrelevant statements concern noise-mitigation, immigration, or law and order. Category 6 is also correct if the statement is simply about judicial fairness, the rule of law, or the right to not be physically harmed.

3. Procedure: Implementing the six different scenarios on CrowdFlower.com

One challenge with designing a crowdcoding task at the moment is that there are few standard procedures and practices. This is due to the lack of relevant studies (political science exceptions are Benoit et al., 2016; Haselmayer and Marcelo 2016; Berinsky et al. 2014, and often draw on advances in computational linguistics). In most cases, our choices reflect the default options and settings on CrowdFlower.com and the recommendations from said studies. In the research on crowd market places at large, there is an ongoing debate about whether higher compensation has positive effects on data quality (Litman et. al. 2015). However, there are a number of other ways in which the results of coding could be improved further: more coders per unit, more minimum time per coding, a higher threshold to pass the quiz, more accomplished/trusted coders than the ones we use (as Peer et al. 2014 recommend), so called dynamic judgements (collecting more judgements for controversial items), a lower limit of maximum codings per person, monetary bonuses and instant feedback for over-performing coders, and a lower threshold for deselecting under-performers). What is uncontroversial is that the validity of codings stands and falls with the quality of the test questions and the respective answers. It is thus important to carefully select and pre-test the test questions. On the one hand, test questions should be clear enough for a serious coder to achieve the 70% standard trust score that coders must achieve in a test-quiz and that they must maintain in a coding job. On the other hand, the test questions must be tough enough to deselect spammers. We have tested 76 potential test questions with 4 university educated (German) persons that had a maximum of 38 minutes (76*30 seconds) for the coding task (2 male, 2 female, 2 social scientists, 2 scientists). On average, 3 out of 4 coders agreed with the coding decision of the experts (74%). This indicates that the test questions are neither too easy nor too hard. We then dropped the 10 statements for which none or only 1 out of 4 test coders (25%) agreed with the expert judgement

(appendix 3 lists the 66 test questions, the 10 questions with to low agreement among the 4 offline coders, and the Horn et al. scores), as the inclusion of such questions also deselects honest and (otherwise) good coders. This left us with 66 test questions; which of course means that the agreement rate is now higher than 74%. A later discussion with each of the four test coders revealed no systematic problems. The 66 test questions were then – together with the correct answers – marked as "Gold" questions in a spreadsheet from which CrowdFlower.com extracts the statements it forwards to online coders. We started the six scenarios at different days and with otherwise completely identical setups. As one of the findings of Benoit et al. (2016) is that the judgements at the statement level already converge strongly from 5 coders on, each statement is coded by at least 5 coders. When the coders achieve 70% accuracy (7 out of 10 right answers) in the quiz mode, they can start with the actual coding and receive 5 US-cents per judgement from then on. Every screen includes our instructions and 5 statements, of which one is always a test question (but the coder does not know which one). To reduce the influence of individual coders, the maximum number of judgements per coder is 310 out of the 1404*5 *trusted* judgements we collect for each scenario (plus the untrusted judgements of the coders who drop below the accuracy score after the quiz mode, because they fail too many secret test questions). As Benoit et al. (2016) rightly point out, crowdcoding is an iterative process in which small errors can lead to useless results. This is why we first ran the coding jobs on a small subsample of statements; to make sure the test questions work. In the next section, we compare the results of the six scenarios, looking mostly at the match between experts and the crowd (at the statement- and the party level), but also at the costs, the duration, and the coders' satisfaction with the clarity of the instructions and the pay they received. All contributors must *at least* be at experience/trust Level 1. Level 2 and 3 coders have higher trust scores, so such a criterion could have biased results in the direction of a high crowd-expert match.

4. Results at the statement level

Although we are ultimately interested in the results and the content validity at the party level, let us first look at the results regarding the expert-crowd match at the level of finalized statements summarized in Table 3. Finalized means that a statement has been allocated to a category based on all trusted codings. Our match measure is simply a percentage score that varies between 0 and 100% of congruence. Other important parameters such as the average confidence score of the codings based on the coders' trust scores, the job costs, the completion time, and the coder satisfaction (with the clarity of instructions and the pay) are also summarized in table 3.

Unsuprisingly, the match is consistently higher for low task complexity and lower for high task complexity. For instance, without any restrictions on the location of the coders, the match for the complex task is 58%, but 85% for the simple task. These match-scores would be higher if a minimum confidence score at the statement level was used to identify and deselect ambivalent codings. Knowing the level of agreement between coders via confidence scores is, after all, one of the advantages of crowdcoding. In that sense, we think that the match-scores we report here are conservative values. Dynamic judgements (i.e., less judgements are collected if coders agree and more if they disagree) or a generally higher number of coders per statement are other factors that could improve the match.

Furthermore, the large difference between the simple and the complex scenario is more modest when the coder expertise is higher due to a switch from non-German to only German IP addresses. The gap between the expert-crowd match for the complex and the simple task decreases from 20% to 12 % then. The match for the complex task improves consistently if we move from the exclusion of German IPs to the mix of German and non-German IPs, to the German IPs only. Granted, even with a focus on German IPs, the match for the complex task remains at a relatively modest 63%.

Higher coder satisfaction and lower costs (due to the lower number of untrusted yet paid judgements) also indicate that the IP restriction may be beneficial for complex coding tasks, even if this restriction in combination with the complex task increases the job completion time from a few hours to a week.

By contrast, to restrict coder characteristics via IP does not seem to be beneficial at all for less complex coding tasks (such as the decision on whether a statement is positively related to equality). A match of 85.3% for the simple task without IP restriction means that in almost 9 out of 10 cases, the crowd arrives at the same conclusion regarding the categorization of a party statement. As table 3 shows, coder satisfaction is high and the costs are similar with and without IP restriction. In sum, these results indicate that contextual knowledge is beneficial for validity only if the task is complex.

Table 3. Expert-crowd match at statement level and other results

Task complexity		Coder expertise (geographical restriction/IP address)		
		Non-German IP	No IP restriction	German IP
High				
	Match:	57.2 %	58.0 %	63.0 %
	Trust:	Ø 71.8	Ø 72.2	Ø 76.1
	Costs:	652.2 $	665 $	538 $
	Time:	3 h	7 h	7 d
	Coder:	3.1 / 5	2.7 / 5	3.3 / 5
Low				
	Match:	87.7 %	85.3 %	74.6 %
	Trust:	Ø 91.0	Ø 88.1	Ø 86.2
	Costs:	527.4 $	533 $	509 $
	Time:	2 h	2 h	7 h
	Coder:	3.8 / 5	3.7 / 5	4.2 / 5

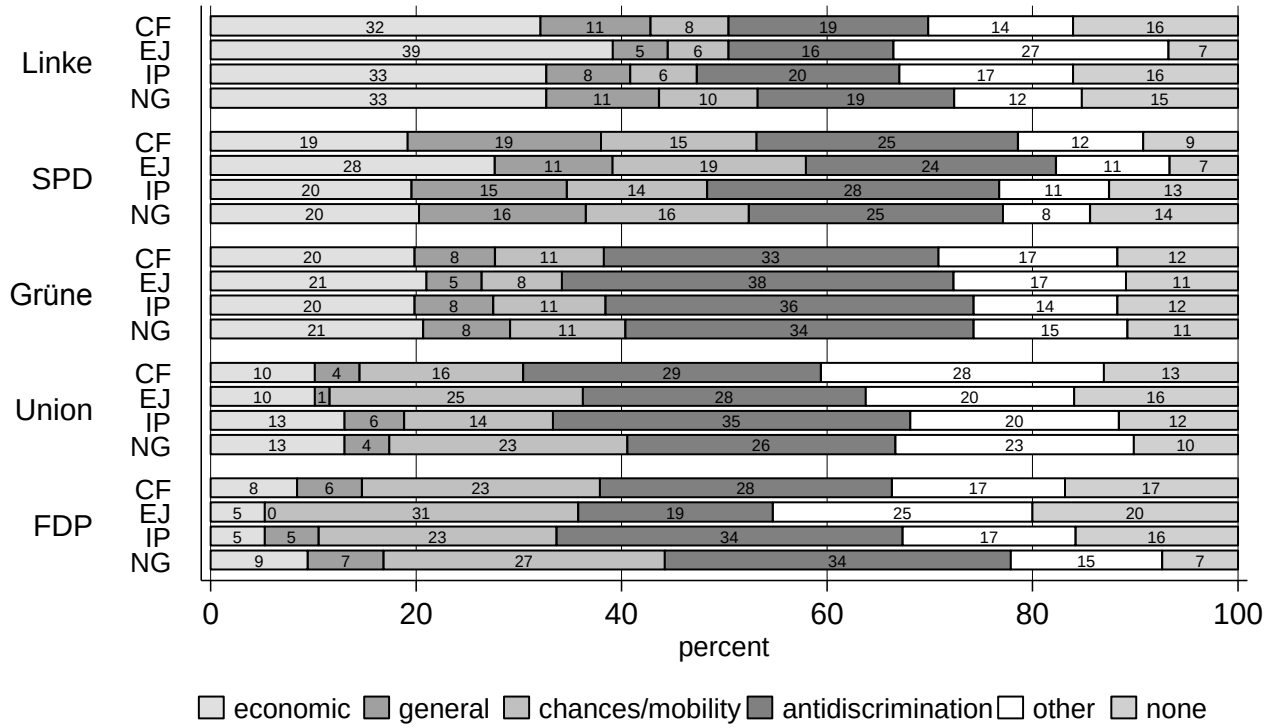
Notes: Match = Expert-crowd match, Trust = CrowdFlower confidence score, Costs = job costs, Time = completion time, Coder = overall coder satisfaction (maximum is 5).

5. Results at the party level

We now turn to the party level. Do the results obtained from crowd-coding – with and without IP restrictions – differ from the results based on expert judgements? We focus on the complex scenarios here because they are less likely to produce valid results – both a priori and in light of the results. These results at the statement level (table 3) confirm that crowdcoding is suitable for simple tasks (in line with Benoit et al., 2016 and Lehman and Zobel, 2017), but the match for the complex task was more modest. Figure 1 maps the emphasis on different concepts of equality for each of the five largest German parties that competed in the last Bundestag-election in 2013. For each row in this stacked bar chart, the shares add up to 100% of positive references to equality (NB: the absolute number of references varies from party to party, e.g. from the Greens with 614, to FDP with 95). The row *CF* shows the distribution according to the crowd coded results. The *EJ*-row shows the distribution based on expert judgments (Horn et al. 2017). The row *IP* shows the emphasis on different equality concepts based on crowdcoding, when access to the coding job is restricted to coders in Germany, while *NG* reverses this restriction and is based on coders outside of Germany. Despite notable deviations, the main conclusions would be the same based on all four rows. The visual inspection reveals that there is a clear left-right gradient when we move from the more left to the more right parties. In all rows, the further we move to the right of the political spectrum, the less emphasis is put on economic equality. Conversely, the centre-right and right parties CDU/CSU and FDP put more emphasis on equal chances. The very strong emphasis the Greens put on antidiscrimination is a third core characteristic of the EJ-results that is confirmed irrespective of whether crowdcoding is conducted using an IP restriction or not. By contrast, and this is an unsurprising yet important qualification, when the absolute number of statements belonging to a

category is very low – as in the case of general mentions of equality (e.g., "We are the party of equality") by the FDP – small errors at the statement level lead to clear deviations at party level.

Figure 1. Distribution of subcategories in the MARPOR equality item based on crowd and experts



Note: CF = based on crowdcoding via CrowdFlower without IP restriction; EJ = Expert Judgments; IP = crowdcoding via CrowdFlower only with coders that have a German IP; NG = crowdcoding via CrowdFlower of coders with non-German IPs. The value labels listed here are rounded values.

Yet, the Pearson's product-moment correlations at the party level between the shares underlying CF, EJ, IP and NG listed in Table 4 are high. Since we are interested in the crowd-expert match, our yardstick are still the expert judgments (EJ). The results gathered using German IPs are closer to the expert judgments (0.93) than the results gathered without IP restriction (0.91) or the reverse restriction (0.89), though these differences in the (significant) correlations in table 4 are not themselves significant (using Stata's *cortesti* module). The scatterplot in appendix 2 shows that

these relationships are linear and not driven by single observations. Overall, the associations at the party level show that results obtained from crowdcoding based on carefully calibrated test questions can match the results of experts even for complex tasks; in particular when IP restrictions are used.

Table 4. Pearson correlation: results from experts and crowdcoding using different IP settings

	EJ (Experts)	CF (Crowd)	IP (German IP)	NG (non-German)
EJ (Experts)	1			
CF (Crowd)	0.9071	1		
IP (German IP)	0.9246	0.9855	1	
NG (non-German)	0.8931	0.9846	0.9801	1

Note: 5 parties and 6 equality categories lead to 30 shares for each of the 4 data gathering strategies.

6. Conclusion

The idea that the best collective decisions are the product of disagreement and contest, of multiple independent judgements rather than of consensus, has long been popular in science, the corporate world, and popular culture. The aim of this paper was to assess if and under what conditions social scientific data collection procedures could benefit from the "wisdom of the crowd", or more specifically the innovative tool of online crowdcoding (which is a specific form of crowd sourcing). In order to further assess the claim by Benoit et al. (2016) that the results of crowdcoding match the results from experts, we have compared the match between expert judgements and crowd-coded judgements at different levels of task complexity and coder expertise; using IP restrictions as a proxy for coders' contextual knowledge. The results are summarized in Table 3 at the level of statements; and in Figure 1 and Table 4 at the level of parties' concepts of equality.

In sum, we mostly agree with the enthusiasm of Benoit et al. (2016). The comparison of expert judgement-based results with crowdcoded results at different levels of task complexity and with and without geographical restrictions indicate that crowdcoding can help researchers to move beyond canonical data-sets that are often inadequate for specific research questions. If researchers carefully design and extensively pre-test filter- and test questions and "invest" in coder "expertise" by restricting access to those most likely most qualified to code the content in a valid way (something that costs time, not money), even elaborate coding schemes suited for fine-grained analyses can be scaled to a large N with the help of the crowd in a fast and affordable manner. As the main domain of crowdsourcing is breaking down big tasks into smaller tasks then outsourced to the crowd, this possibility to use the crowd for complex tasks is the surprising core finding of this research note. Future studies should look more closely into the link between IP restrictions, coder characteristics, and content validity, especially regarding complex coding tasks. We find a better expert-crowd match at the statement- and party level with the IP restriction for the complex task, but admit that the improvements are moderate.

Hopefully, our results will encourage some scholars – in particular those who take issue with the content validity of the rough proxies often used in political science – to start a crowdcoding project. As crowdcoding – for instance in combination with the new Manifesto Corpus data (Merz, et al., 2016) – offers ambitious researchers the chance to upscale theoretically and substantially meaningful category schemes, it could help to soften trade-offs between reliability, generalizability, and content validity in political analysis. While it is obvious that scalability and reliability are strengths of crowdcoding as a data gathering technique, our results show that this must not mean that content validity is compromised. This is good news, as the digitization of the economy, the availability of encompassing (digital) text corpi, and an increase in the number and the diversity of coders are developments that will make crowdcoding ever more attractive in the future.

References

Benoit, K., Conway, D., Lauderdale, B.E., and Laver, M. (2016) 'Crowd-sourced Text Analysis: Reproducible and Agile Production of Political Data'. *American Political Science Review* 110(2): 278-295.

Berinsky, A., Margolis, M., Sances, M. (2014) 'Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys'. *American Journal of Political Science*, 58(3): 739–753.

Gerring, J. (2012, second edition) *Social Science Methodology: A Unified Framework*. Cambridge: Cambridge University Press.

Haselmayer, M. and Marcelo, J. (2016). Sentiment analysis of political communication: combining a dictionary approach with crowdcoding. *Quality & Quantity*, First Online: 21.09.2016.

Horn, A., Kevins, A., Jensen, C., and Van Kersbergen, K. (2017) 'Peeping at the corpus – What is really going on behind the equality and welfare items of the Manifesto project?' *Journal of European Social Policy*, First online: 17.02.2017.

Jensen, C., Van Kersbergen, K. (2016) The politics of inequality. London: Palgrave.

Lehman, P., Zobel, M. (2017) Chancen und Grenzen der Schwarmintelligenz. Werkstattbericht aus einem Crowd-Coding-Projekt. *WZB Democracy and Democratization Blog* 09.03.2017.

Litman, L., Robinson, J., Rosenzweig, C., (2015) The relationship between motivation, monetary compensation, and data quality among US- and India-based workers on Mechanical Turk. *Behavior Research Methods*. 47(2): 519-528.

Merz, N., Regel, S., Lewandowski, J., (2016) The Manifesto Corpus: A new resource for research on political parties and quantitative text analysis. Research & Politics April-June: 1-8.

Sen, A. (1979). Equality of What? The Tanner Lecture on Human Values. Delivered at Stanford University, May 22, 1979.

Peer, E., Vosgerau, J., Acquisti, A., (2014) Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods*. 46(4): 1023-1031.

Surowiecki, J. (2004) 'The wisdom of crowds' New York: Anchor Books.

Appendix 1: German version of the template used to brief the "crowd-coders" before the Quiz:

Complex Task:

Weist diese Aussage aus einem Parteiprogramm einen positiven Bezug zu Gleichheit, sozialer Gerechtigkeit und/oder gleicher Behandlung aller Menschen auf? Wenn JA, welche der Kategorien 1 bis 5 trifft zu? Wenn NEIN, wählen sie bitte Kategorie 6. Bitte lesen sie sich die sechs Kategorien sorgfältig durch! Falls einmal mehrere Kategorien relevant erscheinen sollten, entscheiden Sie welche aus Ihrer Sicht am deutlichsten betont wird.

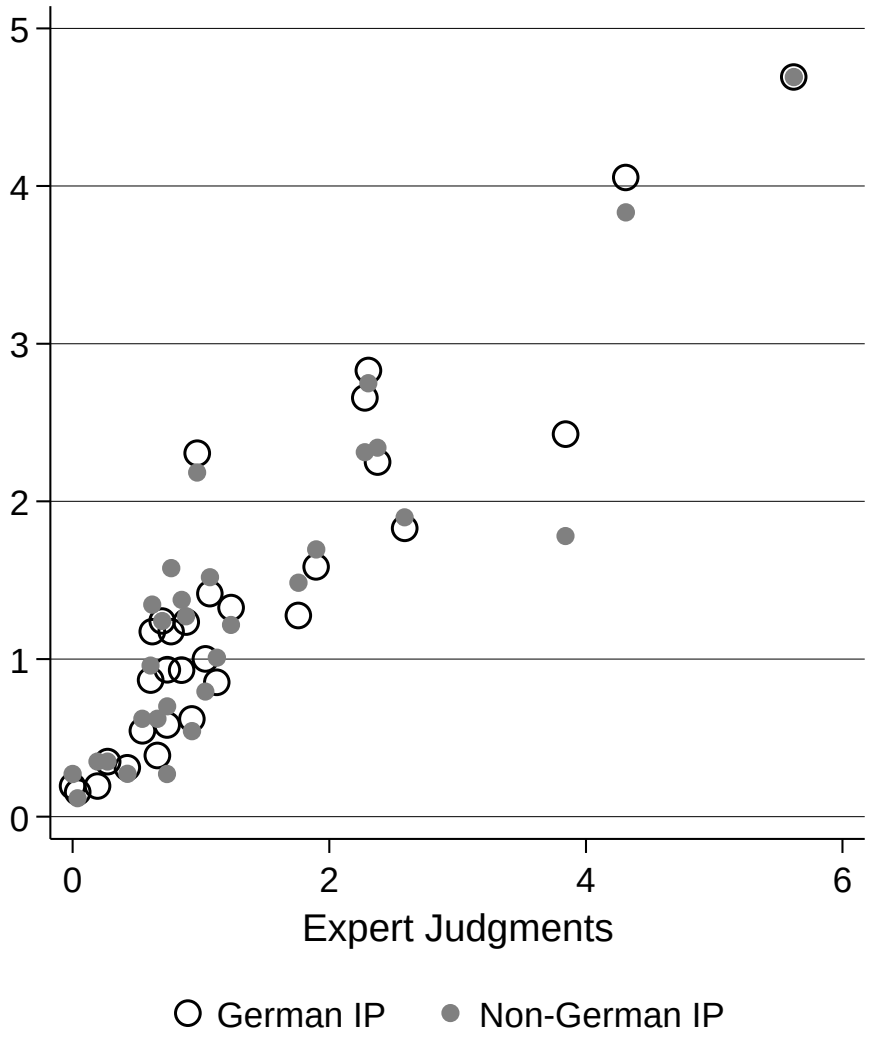
1) **Ökonomische Gleichheit** (Bsp.: wir kritisieren, dass es den kleinen Leuten zu schlecht geht; die wirtschalftiche Ungleichheit in unserem Land hat zugenommen; wir müssen mehr von oben nach unten umverteilen; starke Schultern müssen mehr tragen; die Gesellschaft driftet auseinander; die Schere zwischen arm und Reich muss sich wieder schliessen; Vermögenssteuer jetzt; wir müssen die Gewinne aus der Globalisierung gerecht verteilen; Eigentum für alle bedeuted sozialen Frieden).
2) **Erwähnen** von Gleichheit, (sozialer) Gerechtigkeit, und Solidarität – aber ohne konkret zu werden (Bsp.: wir sind die Partei der Gerechtigkeit; wir stehen für (mehr) Solidarität); soziale Gerechtigkeit).
3) **Chancengleichheit und Soziale Mobilität:** (Bsp.: dass Bildungssystem muss durchlässig sein; mehr Arbeiterkinder müssen es and die Uni schaffen; soziale Herkunft darf nicht über die Zukunft der Kinder entscheiden; jeder muss eine Chance haben – unabhängig vom Geldbeutel der Eltern).
4) **Inklusion, Nichtdiskriminierung, Antidiskriminierung**: (Bsp.: es muss mehr getan werden gegen die Diskriminierung von Frauen, Homosexuellen, Ausländern, Behinderten, Alten; unsere Partei steht für Diversität und Inklusivität; wir wollen eine bunte Gesellschaft; keiner darf diskriminert werden). In Kategorie 4 gehören auch Ausagen zu einer besserer Bezahlung von Frauen (Gender Pay Gap).
5) **Andere:** es gibt einen Bezug zu Gleichheit, (sozialer) Gerechtigkeit und gleicher Behandlung aller Menschen, aber die Aussage passt in keine der vier zuvor genannten Kategorien (Bsp.: globale Gerechtigkeit; Entwicklungshilfe; Gerechtigkeit im Netz; gleicher Zugang zu Internet für alle; Aufbau Ost/der neuen Bundesländer; Mobilität für alle; Energiegerechtigkeit; Klima- und Umweltgerechtigkeit; Wohngerechtigkeit).
6) **NEIN – kein positiver Bezug**. Beispiele fuer irrelevante Ausagen sind Aussagen zu Lärmschutz, Einwanderung, oder Kriminalität. Kategorie 6 ist auch zu wählen, wenn es in der Aussage um die Gleichheit vor dem Gesetz, Rechtsstaatlichkeit, faire Justiz oder körperliche Unversehrtheit geht.

Simple Task:

Weist diese Aussage aus einem Parteiprogramm einen positiven Bezug zu Gleichheit, sozialer Gerechtigkeit und/oder gleicher Behandlung aller Menschen auf?

1) JA
2) NEIN

Appendix 2: Party level results from experts and crowdcoding (with different IP settings)



Appendix 3: Test statements

Experts	**Test statements**
	Statements separated by /, crossed ~~statements~~ were not used since the agreement among the 4 offline-coders was below the 50% threshold.
General	Soziale Gerechtigkeit ist das Programm der LINKEN. /
Economic	Ungleichheit aber kann nur wirksam bekämpfen, wer den Mut hat, Reichtum zu begrenzen und so umzuverteilen, dass er allen zugutekommt. /
Economic	Wir wollen Reiche und Reichtum – Millionäre, Milliardäre, Kapitalvermögen – couragiert besteuern und sicherstellen, dass sie zur Finanzierung des Gemeinwesens angemessen beitragen. /
Economic	und die Sozialpolitik mit einer Umverteilung des Reichtums zu finanzieren, /
Chances	Sie ermöglicht allen – unabhängig von der Herkunft – an Bildung und Entwicklung, überhaupt am gesellschaftlichen Reichtum teilzuhaben. /
General	Beides zusammen bildet die Grundlage einer gerechten und solidarischen Gesellschaft. /
Antidiscr.	Die strukturelle Unterbezahlung von Frauen muss beendet werden. /
None	Dieses Konzept wird in der Partei kontrovers diskutiert. /
General	Eine gerechte Gesellschaft ist für alle besser. /
Other	~~Alle müssen öffentliche Dienste und Einrichtungen nutzen können, unabhängig davon, wo sie wohnen./~~
Chances	Um mehr Bildungschancen für alle zu schaffen, wollen wir das gegliederte Schulsystem überwinden. /

Antidiscr.	Behinderung wird dabei nicht als »Defizit« angesehen, sondern gehört zur menschlichen Vielfalt. /
Antidiscr.	Wir stehen für eine aktive Anti-Diskriminierungspolitik. /
None	Kultur- und Kreativwirtschaft bedarf einer linken Perspektive. /
Other	~~Die Angleichung des Rentenwerts Ost an das Westniveau soll bis Ende 2017 erreicht werden. /~~
None	~~Die Eigentumsverhältnisse an Grund und Boden bilden die Grundlage für ländliche Entwicklung. /~~
Other	Wir wollen einen gleichberechtigten Zugang aller Menschen zum Netz. /
Antidiscr.	Und wir wollen heute etwas ändern, damit wir morgen endlich in einer vielfältigen Gesellschaft leben, in der Kinder, Frauen und Männer, Menschen verschiedener sexueller Identität, verschiedener Religionen, aus unterschiedlichen Kulturen oder unterschiedlicher Herkunft endlich gleichberechtigt leben können und gleiche Möglichkeiten haben./
General	Die große Mehrheit will, dass es in unserer Gesellschaft gerechter zugeht, /
Chances	Gleich ob Schule oder Arbeitsmarkt, Stadtviertel oder Kultureinrichtung, Gesundheit oder Pflege: Unser Ziel ist eine durchlässige Gesellschaft, die Blockaden abbaut, in der Anstrengung belohnt wird und niemand durch verschlossene Türen und gläserne Decken ausgebremst und ausgeschlossen wird. /
Other	Zukunft schaffen – das heißt bezahlbares Wohnen für alle. /
Other	Faire Strompreise erfordern eine gerechtere Finanzierung der Energiewende. /
Economic	Grüne Steuerpolitik ist gerecht, weil stärkere Schultern mehr tragen als schwache./
None	~~Bis dahin ist es aber nicht einzusehen, dass Bund und Länder unterschiedlich hohe Zinsen zahlen müssen – obwohl sie in einer faktischen~~

	~~Haftungsgemeinschaft sind./~~
General	sozialer Ausgleich,/
Chances	Chancengleichheit /
None	~~Wir wollen an die unterschiedlichen Fähigkeiten, Stärken und Qualifikationen dieser Menschen anknüpfen./~~
Antidiscr.	Inklusive Politik fragt nicht nach vermeintlichen Defiziten von Menschen, sondern will die Fähigkeiten der Einzelnen und ihre Teilhabe fördern und unterstützen. /
Antidiscr.	Es ist normal, verschieden zu sein./
Economic	Ziel ist es, die Schere zwischen Arm und Reich zu schließen /
None	~~Rentnerinnen und Rentnern geht es in Deutschland im Durchschnitt noch vergleichsweise gut. /~~
Antidiscr.	Wir wollen Hürden abbauen, damit Menschen in jedem Alter teilhaben können. /
General	Das heißt ein Mehr an sozialer Gerechtigkeit, /
Other	Umweltschutz ist auch eine elementare Gerechtigkeitsfrage und die Voraussetzung für gesellschaftliche Teilhabe. /
Other	Faire Strompreise erfordern eine gerechtere Finanzierung der Energiewende. /
Antidiscr.	Das Recht auf Inklusion muss alle einbeziehen. /
Chances	Es braucht echte Chancengerechtigkeit bei Bildung und Arbeit. /
Antidiscr.	Wir wollen daher einen bundesweiten „Aktionsplan für Vielfalt“, der Homophobie und Transphobie entgegensteuert, /
Antidiscr.	- sorgt für Inklusion und klare Kante gegen Diskriminierungen./

General	für mehr soziale Gerechtigkeit, /
Antidiscr.	Sozialdemokratinnen und Sozialdemokraten stehen seit dem 19. Jahrhundert für die Gleichstellung von Frauen und Männern, /
General	~~Die SPD steht für Gerechtigkeit auch über Generationen hinaus: /~~
None	Auch die Lebensverhältnisse in den ostdeutschen Bundesländern haben sich verbessert. /
Economic	Wir wollen mehr Verteilungsgerechtigkeit bei Einkommen und Vermögen erreichen./
General	und mehr soziale Gerechtigkeit auf den Weg zu bringen /
Economic	Noch nie mussten Vermögende der Gesellschaft, die ihnen den Reichtum ermöglicht hat, so wenig zurückgeben wie heute. /
Other	~~Die schnellere Angleichung der Lebens- und Arbeitsbedingungen in Ost- und Westdeutschland auf der Basis guter Arbeit ist unser Ziel. /~~
General	sozial gerecht und /
None	Fortschritt und Erfolg einer Gesellschaft bemessen sich auch daran, wie Menschen miteinander leben und arbeiten. /
Antidiscr.	„Gender-Mainstreaming" soll wieder durchgängiges Leitprinzip im Regierungshandeln sein./
Chances	Allein die Ziele und Wünsche, der Eifer und die Potenziale der Menschen sollen über Bildungswege entscheiden./
Chances	und Chancengleichheit im Bildungssystem verwirklichen./
None	~~Nicht jeder Mensch fühlt sich dem Geschlecht zugehörig, das bei der Geburt festgestellt wurde. /~~
Economic	Das Ehegattensplitting begünstigt die Einverdienerehe und die Steuerklassenkombination III/V führt zu einer unangemessen hohen monatlichen Belastung des niedrigeren Einkommens./
Economic	Das Familiensplitting lehnen wir ab, weil es nur die Spitzeneinkommen begünstigt./

Chances	Um Chancengleichheit zu erreichen, muss Gebührenfreiheit gelten. /
General	und Gerechtigkeit. /
Economic	Die Schere der Einkommens- und Vermögensverteilung geht auseinander: /
Other	Ziele unserer Entwicklungspolitik sind die Überwindung von Armut und Hunger in der Welt, /
Other	Gerade in den Städten gehört dazu, dass es ausreichend bezahlbare Wohnungen gibt. /
Chances	Wir stehen für eine Politik, die jedem in unserem Land Chancen auf Aufstieg und eine gute Zukunft eröffnet. /
Economic	In der Regierungszeit von Rot-Grün war die Schere zwischen den unteren und oberen Einkommen auseinandergegangen. /
Economic	Starke Schultern müssen mehr tragen als schwache. /
Antidiscr.	und sprechen uns entschieden gegen jede Form der Diskriminierung von Menschen auf Grund ihres Alters aus. /
Other	Auch auf europäischer Ebene setzen wir uns dafür ein, dass der Aufbau Ost weiterhin unterstützt wird./
None	~~Insbesondere müssen Lebensbedingungen und Integration der Menschen in ihren Heimatländern deutlich verbessert werden, sodass sie dort eine Perspektive haben./~~
None	Wir wollen, dass unsere Städte und Regionen auch weiterhin alle Voraussetzungen dafür haben, zum Erfolg unseres Landes beizutragen./
Antidiscr.	Wir wenden uns zugleich entschieden gegen jede Form des Antisemitismus./
General	und Gerechtigkeit./
Chances	Unser Versprechen ist, dass jeder die Chance bekommen soll, seine Träume selbst zu verwirklichen./
None	Überall in Deutschland./

Chances	Wir Liberale wollen Chancen unabhängig von der Herkunft./
Chances	Arbeitsmarktpolitik ist Chancenpolitik, die mehr Menschen den Einstieg in Arbeit ermöglicht – und damit Chancen für das eigene Leben schafft. /
None	Das Lärmsanierungsprogramm für Altstrecken der Bahn werden wir auf hohem Niveau fortsetzen. /
None	Die Anforderungen an die Mobilität der Bürger nehmen weiter zu, beispielsweise um den Arbeitsplatz zu erreichen oder die Lebensqualität in dünn besiedelten Regionen zu sichern. /
Antidiscr.	Wir wenden uns gegen jegliche Diskriminierung aufgrund von Religion, ethnischer Herkunft, Geschlecht, Behinderung, Alter oder sexueller Orientierung. /