Comparing Person- and Process-centric Strategies for Obtaining Quality Data on Amazon Mechanical Turk

Tanushree Mitra C.J. Hutto Eric Gilbert
Georgia Institute of Technology
Atlanta, GA 30308
{tmitra3, cjhutto}@gatech.edu gilbert@cc.gatech.edu

ABSTRACT
In the past half-decade, Amazon Mechanical Turk has radically changed the way many scholars do research. The availability of a massive, distributed, anonymous crowd of individuals willing to perform general human-intelligence micro-tasks for micro-payments is a valuable resource for researchers and practitioners. This paper addresses the challenges of obtaining quality annotations for subjective judgment oriented tasks of varying difficulty. We design and conduct a large, controlled experiment (N=68,000) to measure the efficacy of selected strategies for obtaining high quality data annotations from non-experts. Our results point to the advantages of *person-oriented strategies over process-oriented strategies*. Specifically, we find that screening workers for requisite cognitive aptitudes and providing training in qualitative coding techniques is quite effective, significantly outperforming control and baseline conditions. Interestingly, such strategies can improve coder annotation accuracy above and beyond common benchmark strategies such as Bayesian Truth Serum (BTS).

Author Keywords
Human Computation, Crowd Sourcing, Mechanical Turk, Experimentation, Qualitative Coding, Micro Task

ACM Classification Keywords
H.5.2. Information interfaces and presentation (e.g., HCI) – User Interfaces: Evaluation/methodology; J.4. Social and Behavioral Sciences: Economics, Sociology

INTRODUCTION
The emergence of crowd-sourced micro labor markets like Amazon Mechanical Turk (AMT) is attractive for behavioral and empirical researchers who wish to acquire large-scale independent human judgments, without the burden of intensive recruitment effort or administration costs. Yet acquiring well-measured *high quality* judgments using an online workforce is often seen as a challenge [11,13,32,35,40]. This has led to scholarly work suggesting quality control measures to address the problem of noisy data [8,17,24,35]. Many of these studies have investigated the effectiveness of various quality-control measures as stand-alone intervention strategies on one-off tasks. How do these measures affect quality when working in tandem? What are the challenges faced in acquiring quality results when the difficulty of subjective judgments increase? The present study addresses these questions.

Building on some of the most promising strategies identified by prior work (e.g., [35]), we design and conduct a large empirical study to compare the relative impacts and interactions of 34 intervention strategies. Specifically, we collected and analyzed 68,000 human annotations across more than 280 pairwise statistical comparisons for strategies related to worker screening and selection, interpretive convergence modeling, social motivations, financial incentives, and hybrid combinations. Further, we compare these interactions against a range of representative *subjective judgment-oriented* coding activities of varying difficulty. Our study makes four principal contributions:

- We reveal several intervention strategies which have a substantial positive effect on the quality of data annotations produced by non-experts, regardless of whether correctness is defined by agreement with the most frequent annotation or as agreement with an accepted expert.
- We find that *person-oriented intervention strategies* tend to facilitate high-quality data coding among non-experts. For example, borrowing analogous concepts from the field of Qualitative Data Analysis (QDA) and adapting them for use by a massive, distributed, transient, untrained, anonymous workforce, we find that *prescreening workers for requisite cognitive aptitudes, and providing basic training* in collaborative qualitative coding methods results in better agreement and improved interpretive convergence of non-expert workers.
- We find that person-oriented strategies improve the quality of non-expert data coders above and beyond those achieved via process-oriented strategies like the Bayesian Truth Serum (BTS) technique (c.f., [30,35]).
- Finally, of particular importance for contemporary AMT researchers, we note that while our results show significant improvements in the quality of data annotation tasks over control and baseline conditions, the baseline quality has improved in recent years. In short, compared to the control-level accuracies of just a few years ago [35], AMT is not nearly the "Wild West" that it used to be.

CHI 2015, April 18 - 23, 2015, Seoul, Republic of Korea

ACM 978-1-4503-3145-6/15/04…$15.00
http://dx.doi.org/10.1145/2702123.2702553

BACKGROUND AND RELATED WORK

Data Coding and Annotation

Qualitative Data Analysis (QDA)—that is, systematically analyzing non-numeric data such as interview transcripts, open-ended survey responses, field notes/observations, and a wide range of text documents, images, video, or audio data—is generally a specialized skill most often acquired through formal training. Such skills are costly, both in terms of the financial demand required to obtain the skillset (in undergraduate or graduate school, for example), and in terms of the time, labor, and expense needed to employ the skills. *Qualitative coding*, or the process of interpreting, analyzing, classifying, and labeling qualitative data (e.g., with themes, categories, concepts, observations, attributes or degree anchors, etc.) is a critical step in the larger overall QDA process. As part of qualitative data analysis, many lead researchers employ multiple skilled qualitative *coders* (individuals who perform QDA annotations), each working independently on the same data. Such a strategy makes an explicit trade-off for labor and expense for an increase in accuracy, higher reliability, and a reduction in potential coding errors. What if we could quickly, inexpensively, and yet *reliably* obtain *high-quality* annotations from a massive, distributed, untrained, anonymous, transient labor force?

Crowdsourcing Qualitative Coding & Content Analysis

Crowd-sourced labor markets are an attractive resource for researchers whose studies are conducive to online (Internet-based) participation. Research study data such as qualitative content analysis can be obtained relatively cheaply from potentially thousands of human coders in a very short time. For example, prior work by Wang, Kraut, & Levine [43] asked workers to code discussion forum messages for whether they offered information or provided emotional support. Sorokin & Forsyth [40] had coders annotate images to locate people. Hutto & Gilbert [13] asked coders to annotate the intensity of sentiments in social media texts. Soni and colleagues [39] had workers mark the degree of factuality for statements reported by journalists and bloggers. Andre, Kittur, and Dow [1] asked workers to extract thematic categories for messages shared amid Wikipedians.

Clearly, crowdsourcing does enable quick, inexpensive content analysis and data coding at large scales (c.f., [1,13,39,40,43]). However, these types of QDA activities are often quite subjective in nature. As such, they are susceptible to conflicting interpretations, dissimilar rubrics used for judgments, different levels of (mis)understanding the instructions for the task, or even opportunistic exploitation/gaming to maximize payouts while minimizing effort. Unfortunately, worker anonymity, lack of accountability, inherent transience, and fast cash disbursements can entice the online labor workforce to trade speed for quality [8]. Consequently, the collected annotations may be noisy and poor in quality. Moreover, quality can be inconsistent across different kinds of coding tasks of varying difficulty [35]. Scholars using AMT must therefore carefully consider strategies for ensuring that the codes and annotations produced by non-experts are of *high quality*—that is, ensuring that the coding produced by anonymous workers is accurate and reliable [11,13,32,35,40]. Previous research suggests several quality control measures to tackle the problem of noisy data [1,8,12,17,19,24,35,37]. Most of these earlier works, in isolation, investigate a select set of specialized interventions, often for a single (or just a few kinds of) coding or annotation tasks. Many studies also do not address the challenges associated with coding *subjective judgment* oriented tasks of varying difficulty. To address these gaps, we design and conduct a large empirical study to compare the relative impacts and interactions of numerous intervention strategies (including over 280 pairwise statistical comparisons of strategies related to worker screening and selection, interpretive convergence modeling, social motivation, financial incentives, and hybrid combinations – we discuss these strategies in greater detail later). Further, we compare these interactions against a range of coding activities that have varying degrees of subjective interpretation required.

Crowdsourcing Data Annotations for Machine Learning

Interest in high-quality human annotation is not limited to qualitative method researchers. Machine learning scholars also benefit from access to large-scale, inexpensive, human intelligence for classifying, labeling, interpreting, or otherwise annotating an assorted variety of "training" datasets. Indeed, human-annotated training data acquisition is a fundamental step towards building many learning and prediction models, albeit an expensive and time-consuming step. Here again, the emergence of micro-labor markets has provided a feasible alternative for acquiring large quantities of manual annotations at relatively low cost and within a short period of time—along with several researchers investigating ways to improve the quality of the annotations from inexpert raters [14,36,38]. For example, Snow and colleagues [38] evaluate non-expert annotations for a natural language processing task; they determined how many AMT worker responses were needed to achieve expert-level accuracy. Similarly, Sheng and colleagues [36] showed that using the most commonly selected annotation category from multiple AMT workers as training input to a machine learning classifier improved the classifier's accuracy in over a dozen different data sets. Ipeirotis, Provost, & Wang [15] use more sophisticated algorithms, which account for both per-item classification error and per-worker biases, to help manage data quality subsequent to data annotation.

Whereas these studies concentrate heavily on post-hoc techniques for identifying and filtering out low quality judgments from inexpert coders *subsequent* to data collection, we follow in the same vein as Shaw et al. [35] and focus on *a priori* techniques for encouraging workers to provide attentive, carefully considered responses in the first place. Along with the most promising strategies identified by Shaw et al. [35], we add numerous other person-centered and process-centered strategies for facilitating high quality data coding from non-experts across a range of annotation tasks. We describe these strategies in the next section.

STRATEGIES FOR ELICITING QUALITY DATA

We consider four challenges that affect the quality of crowd annotated data, and discuss strategies to mitigate issues associated with these challenges.

Challenge 1 – Undisclosed cognitive aptitudes

Certain tasks may require workers to have special knowledge, skills or abilities, the lack of which can result in lower quality work despite spending considerable time and effort on a task [16]. As in offline workforces, some workers are better suited for particular tasks than others. Asking anonymous workers with unidentifiable backgrounds to perform activities without first verifying that the worker possesses a required aptitude may result in imprecise or speculative responses, which negatively impacts quality.

Strategy 1 – Screen workers

On AMT, requesters often screen workers from performing certain Human Intelligence Tasks (HITs) unless they meet certain criteria. One very common screening tactic is to restrict participation to workers with an established reputation – e.g., by requiring workers to have already completed a minimum number of HITs (to reduce errors from novices who are unfamiliar with the system or process), and have approval ratings above a certain threshold (e.g., 95%) [2,23,29]. This approach has the benefit of being straightforward and easy for requesters to implement, but it is naive in that it does not explicitly attempt to verify or confirm that a worker actually has the requisite aptitude for performing a given task. For example, a more targeted screening activity (that is tailored more to content analysis coding or linguistic labeling tasks) would be to require workers to have a good understanding of the language of interest, or to require workers to reside in certain countries so that they are more likely to be familiar with localized social norms, customs, and colloquial expressions [13,39].

Challenge 2 – Subjective interpretation disparity

Qualitative content analysis can often be very subjective in nature, and is therefore vulnerable to differences in interpretations, dissimilar rubrics used for judgments, and different levels of (mis)understanding the instructions for the task by unfamiliar, non-expert workers.

Strategy 2 – Provide examples and train workers

Providing examples to introduce workers to a particular coding or annotation task, and modeling or demonstrating the preferred coding/annotation behaviors can help workers establish consistent rubrics (criteria and standards) for judgment decisions [44]. This is analogous to qualitative researchers sharing a common "codebook"—the compilation of codes, their content descriptions and definitions, guidelines for when the codes apply and why, and brief data examples for reference [34]. Along with the examples, requesters on AMT can then require workers to obtain a specific qualification which assesses the degree to which the worker understands how to perform the *task-specific* content analysis annotation or labeling activity. Guiding workers through the process of doing the task trains and calibrates them to the nature of desired responses. This strategy helps improve intercoder/interrater agreement, or *interpretive convergence* – i.e., the degree to which coders agree and remain consistent with their assignment of particular codes to particular data [34].

Challenge 3 – Existing financial incentives are oriented around minimizing time-on-tasks

The micro-labor market environment financially rewards those who work quickly through as many micro-tasks as possible. Consequently, there is little incentive to spend time and effort in providing thoughtfully considered quality responses. If unconsidered judgments and random, arbitrary clicking will pay just as well as thoughtful, carefully considered responses, then some workers may attempt to maximize their earnings while minimizing their effort.

Strategy 3 – Financially incentivize workers to produce high-quality results

In an effort to incentivize carefully considered responses, rewarding high quality responses has shown to improve annotation accuracy [13,35]. For every intervention strategy we examine, we include both a non-incentivized and an incentivized group, and we confirm whether financial incentives continue to have significant impacts above and beyond those of a particular intervention strategy.

Challenge 4 – Low independent (individual) agreement

There are several ways to measure the accuracy of any individual coder. A simple approach is to calculate a percent correct for codes produced by a given coder against an accepted "ground truth." Other useful metrics are Cohen's kappa statistics for nominal coding data and Pearson's correlation for ordinal or interval scales. Regardless of how accuracy is measured, the correctness of any individual coder is often less than perfect due to differences in subjective interpretations.

Strategy 4 – Aggregating, iteratively filtering, or both

One way to mitigate the problem is to use aggregated data, or by searching for congruent responses by taking advantage of the wisdom-of-the-crowd[1] and accepting only the majority agreement from multiple independent workers [13,42]. However, it is often still difficult to obtain meaningful (or at least interpretable) results when aggregated responses are noisy, or when large variance among worker judgments challenge the notion of majority agreement [41]. Prior research has addressed this challenge by adding iterative steps to the basic parallel process of collecting multiple judgments [11, 22]. In other words, use crowd-workers to scrutinize the responses of other workers, thereby allowing human judges (as opposed to statistical or computational processes) to identify the best quality annotations [21,23].

[1] *Wisdom-of-the-crowd* is the process of incorporating aggregated opinions from a collection of individuals to answer a question. The process has been found to be as good as (often better than) estimates from lone individuals, even experts [42].

OUR TASKS

In order to establish a framework of strategies for obtaining high quality labeled data, we administered a combination of the above described strategies across four sets of labeling tasks: identifying the approximate number of people in a picture, sentiment analysis, word intrusion, and credibility assessments (we describe these in more depth in a moment).

Each of annotation task varied in the degree of subjective interpretation required. We deployed four HITs on AMT using a modified version of the NASA-TLX workload inventory scale to assess subjective judgment difficulty [10]. Response options ranged from "Very Low" to "Very High" on a seven-point scale. Each HIT asked 20 workers to perform the four qualitative coding tasks described below, and paid $0.75 per HIT. To account for item effects, we used different content for each annotation task in each of the four HITs. Also, to account for ordering effects, we randomized the order in which the tasks were presented. Thus, we collected a total of 80 responses regarding the difficulty of each type of task, providing us with a range of tasks that vary in their underlying subjective judgment difficulty.

TASK 1: People in Pictures (PP), median difficulty = 1

In this task, we presented workers with an image and asked them to estimate the number of people shown in the picture. This is a well-known data annotation activity in the computer vision research area [7,28]. We selected 50 images containing different numbers of people from the Creative Commons on Flickr[2]. The number of people in each image differed by orders of magnitude, and corresponded to one of five levels: None, About 2 – 7 people, About 20 – 70 people, About 200 – 700 people, and More than 2,000 people.

Expert Annotation / Ground Truth – We determined ground truth at the time we selected the image from Flickr. We purposefully selected images based on a stratified sampling technique such that exactly ten pictures were chosen for each coding/annotation category.

Figure 1: Example pictures for three of the five possible data coding/annotation categories.

TASK 2: Sentiment Analysis (SA), median difficulty = 2

In this task, we mimic a sentiment intensity rating annotation task similar to the one presented in [13] whereby we presented workers with short social media texts (tweets) and asked them to annotate the degree of positive or negative sentiment intensity of the text. We selected 50 random tweets from the public dataset provided by [13]; however, we reduced the range of rating options from nine (a scale from –4 to +4) down to five (a scale from –2 to +2), so that we maintain consistent levels of chance for coding the correct annotations across all our subjective judgment tasks.

Figure 2: Example of the sentiment analysis annotation task

Expert Annotation / Ground Truth – We derived ground truth from the validated "gold standard" public dataset provided by [13], and adjusted by simple binning into a five point annotation scale (rather than the original nine point scale). One of the authors then manually verified each transformed sentiment rating's categorization into one of the five coding/annotation category options.

TASK 3: Word Intrusion (WI), median difficulty = 2

In this task, we mimic a human data annotation task that is devised to measure the semantic cohesiveness of computational topic models [5]. We presented workers with 50 "topics" (lists of words produced by a computational Latent Dirichlet Allocation (LDA) process [3]) created from a collection of 20,000 randomly selected English Wikipedia articles. LDA is a popular unsupervised probabilistic topic modeling technique which originated from the machine learning community. The topics generated by LDA are a set of related words that tend to co-occur in related documents. Following the same procedure described in [5], we inserted an "intruder word" into each of the 50 LDA topics, and asked workers to identify the word that did not belong.

Figure 3: Example of a topic list (with the intruder word highlighted with red text for illustration purposes)

Expert Annotation / Ground Truth – A computational process (rather than a human) selected the intruder word for each topic, making this data annotation task unique among the others in that coders are asked to help establish "ground truth" for the word that least belongs. As such, there was no "expert" other than the LDA computational topic model.

TASK 4: Credibility Assessment (CA), median diff. = 3

In this task, we asked workers to read a tweet, rate its credibility level and provide a reason for their rating. This task aligns with scholarly work done on credibility annotations in social media [4,25,31]. To build a dataset of annotation items that closely resembles real-world information credibility needs, we first ensure that the dataset contains information sharing tweets, specifically those mentioning real world event occurrences [26]. To this end, we borrowed existing computational approaches to filter event specific tweets from the continuous 1% sample of tweets provided by the Twitter Streaming API [4,20,45].

[2] https://www.flickr.com/creativecommons/by-2.0/

Figure 4: Example of a tweet along with the five credibility coding/annotation categories modeled according to existing work on credibility annotation categories [4,39]**.**

Next, we recruited independent human annotators to decide whether a tweet was truly about an event, filtering out false positives in the process. After training the annotators to perform the task, if 8 out of 10 workers agree that a tweet is an event, we add the tweet as a potential candidate for credibility assessment. Next, the first author manually inspected the filtered list to verify the results of the filtering step before sending tweets for credibility assessments on AMT.

Expert Annotation / Ground Truth – Fact-checking services have successfully employed librarians to provide expert information [18]. We recruited three librarians from a large university library as our expert raters. The web interface used to administer the annotation questions to the librarians was similar to the one shown to AMT workers.

CONDUCT OF THE EXPERIMENTS

A full factorial design to evaluate all strategies across all coding/annotation tasks results in combinatorial explosion, making a full factorial experiment intractable. We therefore evaluate the strategies across tasks in stages. A total of 34 combinations were explored (see Table 1). We recruited non-expert content analysis data coders from Amazon Mechanical Turk, and employed a between-group experimental design to ensure we had 40 unique workers in each intervention strategy test condition (i.e., workers were prevented from performing the same data coding activity under different intervention strategies). In each test condition, we asked workers to make coding/annotation decisions for 50 different items (i.e., judgments of the number of people in pictures, sentiments of tweets, intruder words, or credibility assessments). Thus a total of 68,000 annotations were collected (50 items * 40 annotations * 34 intervention strategy combinations).

In the design of our HITs, we leverage insights from [1], who find that presenting workers with context (by having them perform multiple classifications at a time) is highly effective. To ensure workers on an average spend equal time (~ 2-5 minutes) on each HIT independent of task type, a pilot test determined the number of items to fix per HIT.

Comparative measures of correctness

We establish two measures of correctness to judge the quality of annotation in each task: (1) Accuracy compared to crowd (*Worker-to-Crowd*) and (2) Accuracy compared to experts (*Worker-to-Expert*). While the first counts the number of workers who match the most commonly selected response of the crowd (i.e., the mode), the second counts the number of workers who match the mode of experts. We purposely choose mode over other measures of central tendency to establish a strictly conservative comparison metric which can be applied consistently across all comparisons.

Statistical Analysis

For all our experimental conditions we calculate the proportion of correct responses using both metrics, and conduct χ^2 tests of independence to determine whether these proportions differ across experimental conditions. Next, as a post-hoc test, we investigate the cell-wise residuals by performing all possible pairwise comparisons. Because simultaneous comparisons are prone to increased probability of Type 1 error, we apply Bonferroni corrections to counteract the problem of multiple comparisons. Pairwise comparison tests with Bonferroni correction allow researchers to do rigorous post hoc tests following a statistically significant Chi-square omnibus test, while at the same time controlling the familywise error rate [22,33].

EXPERIMENTS

We next present two experiments. Briefly, the first experiment looks at the application of less-complex, *person-centric* a priori strategies on the three easiest subjective judgment tasks. In Experiment 2, we compare the "winner" from Experiment 1 against more complex, *process-oriented* a priori strategies such as BTS, competition, and iteration.

Experiment 1 (Strategies 1-3, Tasks 1-3)

The experimental manipulations we introduce in Experiment 1 consist of variations of intervention strategies 1 through 3, described previously, as well as a control condition that involves no intervention or incentives beyond the payment offered for completing the HIT. We next describe all control and treatment conditions used in Experiment 1.

1. **Control condition, no bonus *(Control NB)***: Workers were presented with simple instructions for completing the data coding/annotation task. No workers were screened, trained, or offered a financial incentive for high-quality annotations. "NB" stands for **N**o **B**onus.
2. **Financial incentive only *(Control Bonus-M):*** Workers were shown the same instructions and data items as the control condition, and were also told that if they closely matched the most commonly selected code/annotation from 39 other workers, they would be given a financial bonus equaling the payment of the HIT (essentially, doubling the pay rate for workers whose deliberated responses matched the wisdom of the crowd majority). "Bonus-M" refers to bonus based on **M**ajority consensus.

3. **Baseline screening *(Baseline NB)***: Screening AMT workers according to their experience and established reputation (e.g., experience with more than 100 HITs and 95% approval ratings) is a common practice among scholars using AMT [1,6,11,27]. We include such a condition as a conservative baseline standard for comparison. Many researchers are concerned with acquiring high quality data coding/annotations, but if intervention strategies like *targeted screening for aptitude* or *task-specific training* do not substantially improve coding quality above such baseline screening techniques, then implementing the more targeted strategies may not be worth the requester's extra effort.
4. **Baseline w/ financial incentive *(Baseline Bonus-M)***: Workers were screened using the same baseline experience and reputation criteria, and were also offered the financial incentive described above for matching the wisdom of the crowd majority.
5. **Targeted screening for aptitude *(Screen Only NB)***: Prior to working on the data annotation HITs, workers were screened for their ability to pass a short standardized English reading comprehension qualification. The qualification presented the prospective worker with a paragraph of text written at an undergraduate college reading-level, and asked five questions to gauge their reading comprehension. Workers had to get 4 of the 5 questions correct to qualify for the annotation HITs.
6. **Targeted screening with financial incentive *(Screen Bonus-M)***: Workers were screened using the same targeted reading comprehension technique, and they were also offered the financial incentive for matching the majority when they performed the HIT.
7. **Task-specific annotation training *(Train Only NB)***: In comments on future work, Andre et al. [1] suggest that future research should investigate the value of training workers for specific QDA coding tasks. Lasecki et al. [19] also advocate training workers on QDA coding prior to performing the work. Therefore, prior to working on our data annotation HITs, workers in this intervention condition were required to pass a qualification which demonstrated (via several examples and descriptions) the *task-specific* coding rubrics and heuristics. We then assessed workers for how well they understood the specific analysis/annotation activity; they had to get 8 of 10 annotations correct to qualify.
8. **Task-specific annotation training with financial incentive *(Train Bonus-M)***: Workers were qualified using the same task-specific demonstration and training techniques, and they were also offered the financial incentive for matching the majority consensus.
9. **Screening and training *(Screen + Train NB)*** – This intervention strategy combined the targeted screening technique with the task-specific training technique (i.e., workers had to pass both qualifications to qualify).
10. **Screening, training, and financial incentive based on majority matching *(Screen + Train + Bonus-M)***: Prior to working on the data annotation HITs, workers had to pass both qualifications, and were also offered the financial incentive for matching the majority.

Table 1 summarizes the control and treatment conditions used for Experiment 1 (described above), and previews the test conditions for Experiment 2 (described later).

Results from Experiment 1

Table 2 shows that intervention strategies have a significant impact on the number of "correct" data annotations produced by non-experts on AMT, regardless of whether "correct" is defined by worker agreement with the most commonly selected annotation code from the crowd, or as agreement with an accepted expert. For example, we see from Table 2 that the χ^2 statistic related to the number of correct annotations when compared to the **crowd** is highly significant: χ^2 (df=9, N= 59,375) = 388.86, $p < 10^{-15}$.

Intervention	Subjective Judgment Tasks: People in Pictures (PP) — *Median Difficulty = 1*				Sentiment Analysis (SA) — *Median Difficulty = 2*				Word Intrusion (WI) — *Median Difficulty = 2*				Credibility Assess (CA) — *Median Difficulty = 3*			
Incentive	NB	Basis of Bonus			NB	Basis of Bonus			NB	Basis of Bonus			NB	Basis of Bonus		
		M	B	C		M	B	C		M	B	C		M	B	C
Control	✓	✓			✓	✓			✓	✓						
Baseline	✓	✓			✓	✓			✓	✓						
Screen	✓	✓			✓	✓			✓	✓						
Train	✓	✓			✓	✓			✓	✓						
Both (Screen+Train)	✓	✓			✓	✓			✓	✓				✓	✓	✓
Iterative Filtering													✓			

Table 1: Combinatorial space of experiments: Four task types varying in median subjective judgment difficulty *(People in Pictures, Sentiment Analysis, Word Intrusion, Credibility Assessment),* **two classes of Incentives** *(NB - No Bonus, Bonus)***, three types of bonus incentive** *(M- Majority Consensus, B – BTS, C – Competition)***, six intervention strategies** *(Control, Baseline, Screen, Train, Both, Iterative Filtering).* **A total of 34 combinations were explored** *(marked ✓).*

Accuracy Metric	Task	df	N	χ^2	p
Worker-to-Crowd	All	9	59,375	388.86	$< 10^{-15}$
Worker-to-Expert	All	9	59,375	149.12	$< 10^{-15}$
Worker-to-Crowd	PP	9	20,000	345.73	$< 10^{-15}$
Worker-to-Expert	PP	9	20,000	46.66	$< 10^{-15}$
Worker-to-Crowd	SA	9	19,675	185.49	$< 10^{-15}$
Worker-to-Expert	SA	9	19,675	160.95	$< 10^{-15}$
Worker-to-Crowd	WI	9	19,700	90.74	$< 10^{-15}$
Worke-to-Expert	WI	9	19,700	59.82	$< 10^{-15}$

Table 2: χ^2 tests of independence for Experiment 1.

Likewise, the χ^2 statistic is also highly significant when comparing worker annotations to an **expert**: χ^2 (*df*=9, *N*= 59,375) = 149.12, $p < 10^{-15}$. Furthermore, Table 2 shows that these significant differences are robust across three diverse types of qualitative data coding/annotation tasks.

After seeing a statistically significant omnibus test, we perform post-hoc analyses of all pairwise comparisons using Bonferroni corrections for a more rigorous alpha criterion. Specifically, there are $\binom{10}{2} = 45$ multiple hypothesis tests, so we test statistical significance with respect to $\alpha = \frac{0.05}{45} = 0.001$ for all paired comparisons. In other words, our between-group experimental study design supports 6 sets of 45 comparisons (i.e., $\binom{10}{2} = 45$ pairs) across 3 tasks and across 2 accuracy metrics, for a total of 45x3x2=270 pairwise comparisons; and for all pairs, p-values must be less than 0.001 in order to be deemed statistically significant.

Figure 5 depicts the percentage of correct annotations in each intervention strategy for each type of coding/annotation task, with indicators of the associated effect sizes for pairs with statistically significant differences.

Experiment 2 – Strategies S3-S4 in Task 4

The experimental manipulations of Experiment 2 are informed by the results from Experiment 1. Referring to the pairwise comparison tests from Experiment 1, we see that screening workers for task-specific aptitude and training them to use a standardized, consistent rubric for subjective judgments improves the quality of annotations. Thus we keep screening and training constant across the conditions of Experiment 2. Our Experiment 2 subjective judgment difficulty is even higher than that of the word intrusion task. Based on these observations, we repeat the ***Screen* + *Train* + *(Bonus-M)*** as a benchmark condition for Experiment 2. As test conditions, we then compare a range of incentive schemes and iterative filtering:

1. **Screening, training, and financial incentive based on majority matching (*Screen* + *Train* + *Bonus-M)***: This condition is same as in Experiment 1 and serves as a benchmark for our second study.
2. **Screening, training, and financial incentive based on Bayesian Truth Serum or BTS (*Screen* + *Train* + *Bonus-B*)**: The effectiveness of using financial incentive schemes based on the Bayesian Truth Serum (BTS) technique is reported by Shaw et al. [35]. BTS asks people to prospectively consider other's responses to improve quality. Thus, in this intervention condition, we ask workers for their own individual responses, but we also ask them to predict the responses of their peers. They were told that their probability of getting a bonus would be higher if they submit answers *that are more surprisingly common* (the same wording as [30,35]).
3. **Screening, training, and financial incentive based on Competition (*Screen* + *Train* + *Bonus-C*)**: In this condition workers are incentivized based on their performance relative to other workers. Workers were told that their response reason pairs will be evaluated by other workers in a subsequent step to determine whether their response is the most plausible in comparison to their peers' responses. They were rewarded when their response was selected as the most plausible.
4. **Screening, training, and Iteration (*Screen* + *Train NB – Iteration)***: This strategy presented workers with the original tweets as well as the response-reason pairs collected in condition 3. Workers were asked to pick the most plausible response-reason pair. Rather than doing credibility assessments directly, workers were acting as judges on the quality of prior assessments, and helping to identify instances where the most commonly selected annotation from the crowd might not be the most accurate/appropriate – that is, they discover whether the crowd has gone astray.

Results from Experiment 2

We compare the proportion of correct response using our two measures of correctness. In Experiment 2, we find no significant difference when using *Worker-to-Expert* metric. Results are significant for *Worker-to-Crowd*: χ^2 (*df*=3, *N*= 7966) = 115.10, $p < 0.008$. To investigate the differences further we again conduct pairwise comparisons with Bonferroni correction. For our four experimental conditions, we conduct a total of $\binom{4}{2} = 6$ comparisons, thus increasing the rigor of our alpha significance criterion to $\alpha = \frac{0.05}{6} = 0.008$.

We find that across all conditions the winning strategy is the one in which workers are screened for cognitive aptitude, trained on task-specific qualitative annotation methods, and offered incentives for matching the majority consensus from the wisdom of the crowd. Surprisingly, comparing the three incentive conditions (majority-based, BTS-based, and competition-based incentives) and the iterative filtering strategy, the BTS strategy is the least effective. There is no significant difference between the effectiveness of competition versus iteration treatments. To summarize the statistical impact of each strategy:

$$S + T + BonusM > [Competition <> Iteration] > BTS$$

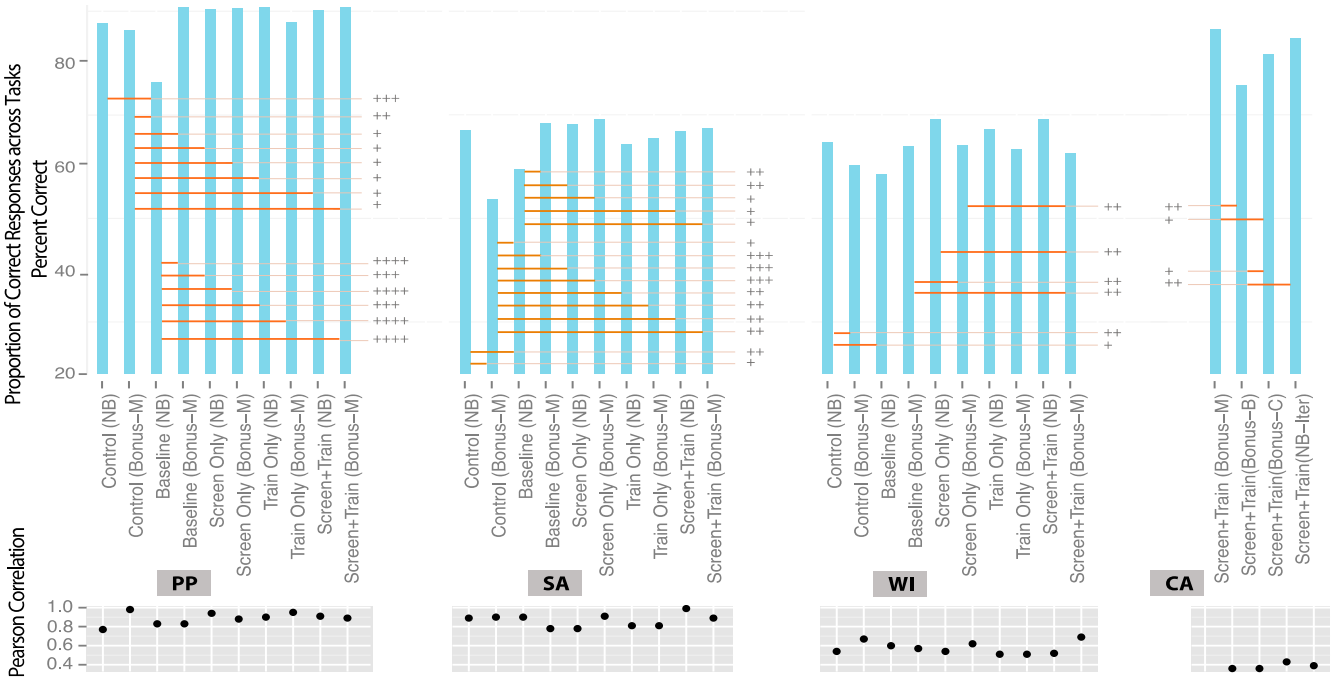

Figure 5: (Top panel) Proportion of correct responses across all tasks with respect to crowd. Pairwise comparisons which are statistically significant are shown with connecting lines (all p-values significant at 0.001 after Bonferroni correction). Effect sizes, as measured by Cramer's V coefficient, are indicated using "+" symbols at four levels: +, ++, +++, and ++++ indicate a very weak effect Cramer's V < 0.15, a weak effect Cramer's V ϵ (0.15, 0.2], a moderate effect (Cramer's V ϵ (0.2, 0.25], and moderately strong Cramer's V ϵ (0.25, 0.3]. (Bottom panel) Pearson correlation between expert and crowd annotations across all tasks.

DISCUSSION AND CONCLUSIONS

We systematically compared the relative impacts of numerous a priori strategies for improving the quality of annotations from non-experts, and we checked their robustness across a variety of different content analysis coding and data annotation tasks. We offer several reasons for focusing on a priori techniques, as opposed to complex statistical data cleaning techniques performed post-collection. First, the value of a priori strategies are not as well explored, lending novelty to our contributions. Second, a priori person-oriented strategies emulate the procedures of sharing a common QDA codebook. Our results demonstrate the value of applying a well-established method for qualitative data coding to crowd-sourced data annotations by non-experts. Third, screening and training techniques have a onetime upfront cost which soon amortizes with increases in the size of datasets, so the techniques scale exceptionally well. Fourth, person-oriented strategies are arguably more generalizable; they can be adapted to adjudicate both *objective* and *subjective* judgments. Post hoc data cleaning is suited more for objective tasks and breaks down as data becomes noisy; thus, post hoc procedures are of limited use for subjective oriented judgment tasks. Fifth, for time sensitive judgments (e.g., credibility decisions for rapidly unfolding events), simple a priori methods trump complex post hoc methods.

Crowd generated data annotations by non-experts can be reliable and of high-quality

We find that our crowd-generated data annotations have relatively high data quality (in comparison to prior research, e.g., [35]), even though we use aggressive criteria for measuring accuracy; that is, we purposely choose *exact* matching with the mode over other potential measures (e.g., mean or median) as a strict metric for all comparisons. Further, the effects of interventions are generally robust across a range of representative QDA data annotation tasks of varying difficulty and with varying degrees of subjective interpretation required. For example, the top panel of Figure 5 shows the agreement between individual coders and the crowd provided ground truth. In every task, the agreement is well above chance (20% for all tasks).

As data coding tasks become more subjectively difficult for non-experts, it gets harder to achieve interpretive convergence. This is demonstrated by the decreasing correlation trend for both the top and the bottom panels in Figure 5. When compared to an accepted expert (Figure 5, bottom), we find generally high agreement between experts and the crowd-produced accuracy measure for the two easier subjective judgment tasks (i.e., judging the number of People in Pictures [PP] and Sentiment Analysis [SA] for tweets). The subjective judgment difficulty of the Credibility Assessment (CA) task is quite high, and so correlation of the crowd to the expert librarians is understandably decreased, though still moderately strong in the 0.4 to 0.45 range. (Note: the lower correlation for the Word Intrusion [WI] task is related more to the poor performance of the computational topic model algorithm as a non-human "expert" than the ability of the crowd to match that expert).

Person-oriented strategies trump process-oriented strategies for encouraging high-quality data coding

In general, we find that screening and training workers are successful strategies for improving data annotation quality. Financial incentives do not appear to help improve quality, except in the simplest baseline condition (impact becomes negligible when stronger person-centric strategies are used).

The insightful work from Shaw and colleagues [35] noted that process-centric strategies like BTS were effective at promoting better quality annotations. An interesting finding from this study is that (in contrast to commonly employed process oriented tactics) when we target intervention strategies towards verifying or changing specific attributes of the individual worker, we see better and more consistent improvements in data annotation quality. By verifying that a person has the requisite cognitive aptitude (knowledge, skill, or ability) necessary to perform a particular qualitative data annotation task, together with training workers on qualitative data coding expectations, we can significantly improve effectiveness *above and beyond* the effects of BTS (see Figure 5, top). Person-centric strategies, such as a prior screening for requisite aptitudes and prior training on task specific coding rules and heuristics, emulate the processes of personnel selection and sharing a common "codebook". Qualitative scholars have been using such strategies for years to facilitate accurate and reliable data annotations among collaborative data coders [34]. By applying these techniques to crowd-sourced non-expert workers, researchers are more likely to achieve greater degrees of interpretive convergence – and do so more quickly, with less variation (c.f., [9]) – because workers are thinking about the data coding activity in the same ways.

Why do more to get less?

In terms of effort on behalf of both the research-requester and the worker-coder, intervention strategies such as screening and training workers have a one time up-front cost associated with their implementation, but their cost quickly becomes amortized for even moderate sized datasets. In contrast, strategies such as BTS, Competition, and Iteration require the same, sustained level of effort for every data item that needs to be coded or annotated. As such, the per-item cost for BTS, Competition, and Iteration are much heavier as the size of the dataset grows. Given that these more complex strategies actually do not perform as well as screening and training, why do more to get less quality?

Amazon is not a neutral observer; AMT is getting better

While our results show better consistency in the quality of data annotation tasks when using person-centric strategies over control and baseline conditions, we note that the quality of data obtained from AMT workers in those conditions is much higher than we initially expected, given our experience with the platform over the years. We highlight the finding that in every task across all interventions, the accuracy of crowd-produced annotation is not only well above random chance (20% for all tasks), but also well above the control condition and even the BTS treatment condition for similar subjective-oriented tasks reported in [35] just a few years ago. For example the "rank content" and "rank users" tasks from [35] are precisely the kind of subjective-oriented tasks that we are targeting with our person-centric interventions, but accuracy reported in [35] peaks at ~40% for even the best incentive category (BTS). (Note: the chance for randomly guessing the correct response for these two subjective judgment tasks was also 20%, the same as with our study). Contrast this with our results; even in the more difficult subjective judgment tasks (SA, WI, and CA), we find control condition accuracies in the 55-80% range (and our person-oriented treatment conditions are even higher, in the 65-90% range). These performance scores far exceed those reported in [35] for the two subjective judgment tasks. So it seems that compared to just a few years ago, AMT is not nearly the "Wild West" that it used to be. Interestingly, the kinds of quality control measures Amazon has enacted are also quite person-centric: e.g., requiring workers to verify their identity by providing their tax information[3], requiring workers to prove their humanity using CAPTCHAs at random intervals before accepting some HITs, and so on. (The first author even received a request from AMT to provide proof of U.S. residence by faxing a utility bill).

Our results are not intended necessarily to be prescriptive. Even in a study this size, we still focus on just a subset of potential intervention strategies, subjective judgment tasks, and various social and financial based incentives. Future work should directly compare the efficiency and effectiveness of a priori person-centric techniques to peer-centric methods (c.f., [12]) and more complex post hoc statistical consensus finding techniques (c.f., [37]). Nonetheless, the person-centric results reported in this paper help illustrate the value of applying established qualitative data analysis methods to crowd-sourced QDA coding by non-experts.

REFERENCES

1. Andre, P., Kittur, A., and Dow, S.P. Crowd synthesis: extracting categories and clusters from complex data. *CSCW* 2014, 989–998.
2. Berinsky, A.J., Huber, G.A., and Lenz, G.S. Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. *Political Analysis 20*, 3 2012, 351–368.
3. Blei, D., Ng, A., and Jordan, M. Latent dirichlet allocation. *J. of Machine Learning Res.3*, 2003, 993–1022.
4. Castillo, C., Mendoza, M., and Poblete, B. Information credibility on twitter. *CHI* 2011, 675–684.
5. Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., and Blei, D. Reading Tea Leaves: How Humans Interpret Topic Models. *NIPS* 2009, 288–296.

[3] https://www.mturk.com/mturk/help?helpPage=worker

6. Crump, M.J.C., McDonnell, J.V., and Gureckis, T.M. Evaluating Amazon's Mechanical Turk as a Tool for Experimental Beh. Res. *PLoS ONE 8*, 3 2013.
7. Dalal, N. and Triggs, B. Histograms of oriented gradients for human detection. *CVPR* 2005, 886–893.
8. Downs, J.S., Holbrook, M.B., Sheng, S., and Cranor, L.F. Are your participants gaming the system?: screening mechanical turk workers. *CHI* 2010, 2399–2402.
9. Gilbert, E. What if we ask a different question?: social inferences create product ratings faster. *CHI* 2014.
10. Hart, S.G. and Staveland, L.E. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Adv. in psych. 52*, 1988, 139–183.
11. Heer, J. and Bostock, M. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. *CHI* 2010, 203–212.
12. Huang, S.-W. and Fu, W.-T. Don't hide in the crowd!: increasing social transparency between peer workers improves crowdsourcing outcomes. *CHI* 2013.
13. Hutto, C.J. and Gilbert, E. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *ICWSM* 2014, 216–255.
14. Ipeirotis, P.G. Analyzing the amazon mechanical turk marketplace. *XRDS: Crossroads17*, 2 2010, 16–21.
15. Ipeirotis, P.G., Provost, F., and Wang, J. Quality management on Amazon Mechanical Turk. *HCOMP* 2010.
16. Kazai, G., Kamps, J., and Milic-Frayling, N. Worker types and personality traits in crowdsourcing relevance labels. *CIKM* 2011, 1941–1944.
17. Kittur, A., Chi, E.H., and Suh, B. Crowdsourcing user studies with Mechanical Turk. *CHI* 2008, 453–456.
18. Kriplean, T., Bonnar, C., Borning, A., Kinney, B., and Gill, B. Integrating on-demand fact-checking with public dialogue. *CSCW* 2014, 1188–1199.
19. Lasecki, W.S., Gordon, M., Koutra, D., Jung, M.F., Dow, S.P., and Bigham, J.P. Glance: rapidly coding behavioral video with the crowd. *UIST* 2014, 551–562.
20. Lau, J.H., Collier, N., and Baldwin, T. On-line Trend Analysis with Topic Models: #twitter Trends Detection Topic Model Online. *COLING* 2012, 1519–1534.
21. Little, G., Chilton, L.B., Goldman, M., and Miller, R.C. Exploring iterative and parallel human computation processes. *CHI* 2010, 68–76.
22. MacDonald, P.L. and Gardner, R.C. Type I error rate comparisons of post hoc procedures for I j Chi-Square tables. *Ed. and Psych. Meas. 60*, 5 2000, 735–754.
23. Mason, W. and Suri, S. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior research methods 44*, 1 2012, 1–23.
24. Mason, W. and Watts, D.J. Financial incentives and the performance of crowds. *ACM SIGKDD Explorations Newsletter 11*, 2 (2010), 100–108.
25. Morris, M.R., Counts, S., Roseway, A., Hoff, A., and Schwarz, J. Tweeting is believing?: understanding microblog credibility perceptions. *CSCW* 2012, 441–450.
26. Naaman, M., Boase, J., and Lai, C.-H. Is it really about me?: message content in social awareness streams. *CSCW* 2010, 189–192.
27. Paolacci, G., Chandler, J., and Ipeirotis, P.G. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making 5*, 5 2010, 411–419.
28. Papageorgiou, C. and Poggio, T. A trainable system for object detection. *Int. J. of Computer Vision 38*, 1 2000.
29. Peer, E., Vosgerau, J., and Acquisti, A. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods*, 2013, 1–9.
30. Prelec, D. A Bayesian truth serum for subjective data. *Science 306*, 5695 2004, 462–466.
31. Qazvinian, V., Rosengren, E., Radev, D.R., and Mei, Q. Rumor has it: Identifying misinformation in microblogs. *EMNLP*2011, 1589–1599.
32. Rashtchian, C., Young, P., Hodosh, M., and Hockenmaier, J. Collecting image annotations using Amazon's Mechanical Turk. *In Proc. NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 139–147, 2010.
33. Rosenthal, R. and Rubin, D.B. Multiple contrasts and ordered Bonferroni procedures. *J. Ed. Psy. 76*, 6 1984.
34. Saldana, J. *The Coding Manual for Qualitative Researchers.* SAGE Publications, 2009.
35. Shaw, A.D., Horton, J.J., and Chen, D.L. Designing incentives for inexpert human raters. *CHI* 2011.
36. Sheng, V.S., Provost, F., and Ipeirotis, P.G. Get another label? improving data quality and data mining using multiple, noisy labelers. *KDD* 2008, 614–622.
37. Sheshadri, A. and Lease, M. SQUARE: A Benchmark for Research on Computing Crowd Consensus. *HCOMP* 2013, 156–164.
38. Snow, R., O'Connor, B., Jurafsky, D., and Ng, A.Y. Cheap and Fast - But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. *EMNLP* 2008.
39. Soni, S., Mitra, T., Gilbert, E., and Eisenstein, J. Modeling Factuality Judgments in Social Media Text. *In Proc. ACL*, 415–420, 2014.
40. Sorokin, A. and Forsyth, D. Utility data annotation with Amazon Mechanical Turk. *CVPRW* 2008, 1–8.
41. Sun, Y.-A., Roy, S., and Little, G. Beyond Independent Agreement: A Tournament Selection Approach for Quality Assurance of Human Computation Tasks. *Human Computation*, 2011.
42. Surowiecki, J. *The Wisdom of Crowds.* Anchor Books, New York, NY, 2004.
43. Wang, Y.-C., Kraut, R., and Levine, J.M. To stay or leave?: the relationship of emotional and informational support to commitment in online health support groups. *CSCW* 2012, 833–842.
44. Willett, W., Heer, J., and Agrawala, M. Strategies for crowdsourcing social data analysis. *CHI* 2012.
45. Zhao, W.X., Jiang, J., Weng, J., et al. Comparing twitter and traditional media using topic models. *Advances in Information Retrieval*. Springer, 2011, 338–349.