Testing Conventional Wisdom (of the Crowd)

Noah Burrell[1] **Grant Schoenebeck**[1]

[1]University of Michigan, Ann Arbor, Michigan, USA

Abstract

Do common assumptions about the way that crowd workers make mistakes in microtask (labeling) applications manifest in real crowdsourcing data? Prior work only addresses this question indirectly. Instead, it primarily focuses on designing new label aggregation algorithms, seeming to imply that better performance justifies any additional assumptions. However, empirical evidence in past instances has raised significant challenges to common assumptions. We continue this line of work, using crowdsourcing data itself as directly as possible to interrogate several basic assumptions about workers and tasks. We find strong evidence that the assumption that workers respond correctly to each task with a constant probability, which is common in theoretical work, is implausible in real data. We also illustrate how heterogeneity among tasks and workers can take different forms, which have different implications for the design and evaluation of label aggregation algorithms.

1 INTRODUCTION

As a whole, crowds can be surprisingly wise [Surowiecki, 2004], but individuals within the crowd are unsurprisingly prone to making mistakes. This is true for complex applications like prediction markets—where individuals collaborate to create a probabilistic forecast for some event by trading securities linked to particular outcomes of that event. And it is equally true for simpler applications like microtasks—where individuals collaborate to complete a simple task, like labeling an image, by performing the task individually and submitting their responses to be algorithmically aggregated in a way that will discern the correct label from the various responses with high probability. Given this fallibility, then, it is important to understand how workers make mistakes.

Clearly, how workers make mistakes on individual tasks has important implications for the design of aggregation algorithms. These algorithms frequently leverage insights that flow from assumptions about how errors are made. For example, some algorithms, like those proposed by Burnap et al. [2015] and Welinder and Perona [2010] rely on the assumption that there are expert workers, who label more accurately than a typical worker. Under this assumption, a clear aggregation strategy emerges: identify the experts, then take the majority answer from the expert labels.

But how workers make mistakes on individual tasks also has important implications for the *evaluation* of aggregation algorithms. When an algorithm is tested on a group of data sets, the degree to which the results of those tests will be indicative of performance on future data depends on the degree to which the test data sets are representative of the future data. The (approximate) validity of basic assumptions about how workers make errors are important dimensions along which test data may or may not be representative of future data, because those assumptions are a key factor in the design of aggregation algorithms.

Further, it is important to understand the nature of individual mistakes in order to quantify uncertainty about labels. Uncertainty quantification can be used, e.g., as in active learning, to train classifiers that achieve a given level of accuracy at a lower cost than otherwise [Cerquides et al., 2021, Passonneau and Carpenter, 2014, Sheng et al., 2008]. More specifically, the estimated parameters of an error model can also be used to improve the utility of a labeled data set as a training set for a machine learning algorithm, as shown by Lalor et al. [2019], and to improve the reliability of crowdsourced experiments, as shown by Katsuno et al. [2019].

The importance of understanding errors, thus, leads to a natural question: What evidence is there in real crowdsourcing data for the various assumptions that underlie the common error models? Surprisingly, work towards answering this

The code for our experiments is at `https://github.com/burrelln/Testing-Conventional-Wisdom`.

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

question is largely absent from the literature. The most direct evidence for the utility of an error model is typically just that an algorithm based on it outperforms previous algorithms in aggregating labels to recover the ground truth. However, there are many factors that confound the relationship between an underlying model and the performance of an algorithm that is designed using that model, including the choice of test data sets. As a result, assessing the validity of modeling assumptions solely through algorithmic performance offers a limited view. It is insufficient for supporting any theoretical claims that follow from those assumptions. Moreover, it is insufficient for understanding how representative a group of test data sets is of future data and for assessing the utility of the various error models for uncertainty quantification.

We address these limitations by directly exploring the degree to which there exists evidence in real data sets to justify several common assumptions about worker errors. In doing so, we uncover some regularities that hold across a diverse collection of data sets alongside much variation in other fundamental characteristics. Specifically, we find that:

- Errors depend on the category of each task's correct response. In particular, workers do not have a constant probability of correctness for all tasks (Section 4).
- Whether errors appear to depend on factors beyond the correct response category varies (Sections 5.1 and 5.2).
- Worker proficiency distributions are well-characterized by (generally reliable) modal workers (Section 5.2).
- Exceptionally reliable "expert" workers do not appear to play a significant role (Section 5.2.2).

1.1 RELATED WORK

Challenging Assumptions. In this work, we test common assumptions about crowdsourcing workers and tasks using publicly-available data to help guide future research and practice. Prior work includes three particularly notable examples that do exactly that. Yin et al. [2016] uncover a robust network of communication among crowdsourcing workers. A key lesson of their analysis is that the group of workers that complete a task on Amazon Mechanical Turk (a popular crowdsourcing platform) is *not* a random sample of all active workers, because workers talk to their peers about tasks that are enjoyable, lucrative, etc. More directly in the domain of label aggregation, Li et al. [2019] demonstrate that the common assumption that the average number of labels per task is small (≤ 3) often does not hold. Rather, in the publicly-available data sets they identify, the average number of labels per task is commonly at least 5. Further they show that, in this label-dense setting, many state-of-the-art aggregation algorithms perform no better than a simple majority vote, despite being much more computationally expensive. Lastly, Wei et al. [2022] argue that certain noise models from the image classification literature fail to adequately describe real-world noise in two new benchmark data sets that they introduce for image classification tasks.

In testing the assumptions that we consider, we rely crucially on an additional assumption that is typical in crowdsourcing—that the underlying tasks have an objective ground truth category that can be recovered by aggregating labels. This assumption makes sense for the tasks that we consider, which are relatively simple to complete and in many cases have objectively correct responses. However, it is not an appropriate assumption for all crowdsourcing tasks. Recent work, e.g., by Basile et al. [2021],Gordon et al. [2021], and Plank [2022], has explored alternative approaches to working with crowdsourced data in those settings where it does not make sense to assume that tasks have an objective ground truth.

Aggregating Labels. The assumptions that we focus on in this work are generally implicit in the error models that are employed in the design of label aggregation algorithms. Just two such families of error models are nearly ubiquitous in the literature and they imply different assumptions about the heterogeneity of tasks and workers:

1. Dawid-Skene models (Section 2.1) assume that a worker's errors on a task depend primarily on the correct response for the task and their own proficiency [Dawid and Skene, 1979, Karger et al., 2011, Liu and Wang, 2012, Liu et al., 2012, Raykar et al., 2010, Welinder and Perona, 2010, Zhang et al., 2014].
2. Item response theory models (Section 2.2) generally assume that tasks, independently of the correct response, follow a particular pattern of heterogeneity that affects a given worker's probability of responding correctly in specific ways [Bachrach et al., 2012, Khattak et al., 2016, Whitehill et al., 2009, Welinder et al., 2010][1].

The most common approach in designing a label aggregation algorithm is to adopt a model from one of these families. But there are a few prominent exceptions to this trend. For example, Zhou et al. [2012, 2015] adopt a very flexible error model under which each task-worker pair is associated with its own distribution over the possible responses. This model makes relatively few assumptions about the nature of workers and tasks, but as a result, also has limited utility for extrapolating to unseen examples. Another unique approach by Jung and Lease [2012] is to apply probabilistic matrix factorization, a collaborative filtering technique, to predict labels from each worker on all tasks before aggregating the predicted labels via majority vote or some other algorithm.

Bayesian Annotation Models. In this work, we seek to validate or invalidate assumptions using data itself, and

[1] Otani et al. [2016] also use a model based on item response theory in the related setting of pairwise comparisons.

therefore to be agnostic to specific aggregation algorithms or models, as much as possible. However, certain outcomes are difficult to distinguish without an underlying model. For example, is a certain group of workers with high accuracy a group of experts or were they just assigned easy tasks? As a result, in Section 5.2.1 we fit models from the Dawid-Skene and item response theory families and use the parameters of the best-fitting models to further explore the data. In that section, our work resembles that of Paun et al. [2018] and Lakkaraju et al. [2015], who each consider sets of Bayesian annotation models and evaluate their utility for various estimation and prediction tasks, including label aggregation.

Although Paun et al. and Lakkaraju et al. draw from essentially the same families of models that we consider, our work has significant methodological differences. And, ultimately, we apply models toward a different end—to answer specific questions about the data themselves and how they relate to common assumptions from the label aggregation literature, rather then to answer questions about the relative utility of the models for solving problems where the data is an input or qualitative questions about interpreting the parameters of more complex models.

2 MODELING

The fundamental elements of an error model are *tasks* and *workers*. In our setting, a task is an object that is associated with a collection of *categories* or labels with a fixed, finite size k. Among these categories, exactly one applies to the object. That category is called the *ground truth*. For example, a task might be an image of a duck, with the categories "Duck" and "'No Duck." In that case, "Duck" is the ground truth. A worker's job is to select the ground truth category that applies to the object for each task assigned to them.

2.1 DAWID-SKENE

The most popular models in the label aggregation literature are Dawid-Skene (DS) models, which were proposed as a way to understand and mitigate individual errors in clinical diagnoses [Dawid and Skene, 1979]. In a DS model, the interactions between workers and tasks are parameterized by a collection of *confusion matrices*. A confusion matrix M is a $k \times k$ stochastic matrix. Entry m_{ij} denotes the probability of a worker reporting category j on a task for which the ground truth is i. Typically, each worker is associated with their own confusion matrix, but variants of that basic model include models where a single confusion matrix is shared among a cluster of workers or among the entire population.

Intuitively, DS models suppose that the probability of a particular worker making an error on a particular task can depend on that task's ground truth category, but only on that. In our running example, that means that duck images and non-duck images may have different patterns of errors, but every image of a duck (and every image that is not of a duck) is more or less equally recognizable as such. We decompose this into two distinct assumptions: The first is that the pattern of errors is *category-dependent*. The second is that tasks with the same ground truth are *homogeneous*.

2.2 ITEM RESPONSE THEORY

Item response theory (IRT) [Embretson and Reise, 2000, Reckase, 2009] was developed in psychometrics for the purpose of designing tests (e.g., academic assessments) and interpreting their results. In contrast to DS models, IRT models parameterize both workers *and* tasks. Each worker i is characterized by an *ability* parameter θ_i, which may be a scalar or vector. θ_i represents i's adeptness at the underlying skill being "measured" by a particular test, i.e., set of tasks. Each task j is characterized by up to three scalar parameters: a discrimination parameter a_j, a difficulty parameter b_j, and a "guessing" parameter $c_j \in [0, 1]$. The worker and task parameters interact in the following way to determine the probability of a correct response on task j from worker i:

$$\Pr[\text{correct}] = c_j + (1 - c_j)\,\text{expit}\left(a_j\left(\theta_i - b_j\right)\right), \quad (1)$$

where $\text{expit}(x) = \frac{\exp(x)}{1+\exp(x)}$ is the standard inverse-logit (i.e., logistic) function. In this function, *ceteris paribus*, the discrimination parameter controls the rate of change in the probability of correctness as the worker's ability varies. The guessing parameter captures the intuition that, with a finite number of categories, it is possible to produce the correct response without identifying it by responding randomly.

Equation (1) captures the three basic IRT models. The difference between these models is that the simpler models impose stronger constraints on the task parameters. In the one-parameter logistic model (1PL), $a_j = a$, a constant, and $c_j = 0$ for all tasks j. In the two-parameter logistic model (2PL), $c_j = 0$ for all tasks j. In the three-parameter logistic model (3PL), all of the task parameters are allowed to vary.

On the whole, IRT models suggest that the probability of a particular worker making an error on a particular task comes down to an interaction between the characteristics of the worker and the characteristics of the task. However, the characteristics of tasks that are assumed to be relevant are different under IRT than under DS. In particular, the ground truth category is generally not taken into consideration.

3 DATA

Given a worker and a task, the DS model predicts a complete response distribution over the possible categories. IRT predicts a probability that the response will be correct, but does not predict which category will be chosen if the response is

not correct. To remove this asymmetry, we limit our analysis to data sets with binary categories, so that specifying a probability of correctness is equivalent to specifying a complete distribution over the categories. In addition to binary categories, we also require the data sets to have ground truth labels. This introduces an important assumption we make throughout: that ground truth labels are sufficiently reliable.

Sheshadri and Lease [2013]'s Statistical QUality Assurance Robustness Evaluation ([SQUARE, Accessed: Oct. 2022]) project, a benchmarking resource for label aggregation research, provides six data sets that meet our criteria[2] and are meant to be used to evaluate label aggregation algorithms:

- **BM** involves labeling the sentiment of tweets as either positive or negative [Mozafari et al., 2012, 2014].
- **HCB** involves determining whether a particular Web page is relevant to a given search query [SQUARE, Accessed: Oct. 2022].
- **RTE** involves *textual entailment*, i.e., deciding whether a given statement implies a subsequent one [Snow et al., 2008].
- **TEMP** involves deciding two events' temporal order [Snow et al., 2008].
- **WB** involves labeling images by whether they contain a certain kind of bird Welinder and Perona [2010].
- **WVSCM** involves labeling whether images of smiles are of "Duchenne" smiles [Whitehill et al., 2009].

Several of these data sets are no longer available through the SQUARE project site, so we provide alternative links [Mozafari et al., Snow et al., Welinder et al.].

We also consider the following additional data set, which was released subsequently to the SQUARE project:

- **SP** involves labeling the sentiment of a sentence extracted from a movie review as either positive or negative [Venanzi et al., 2015].

4 CATEGORY-DEPENDENT ERRORS

We begin with a model-agnostic, non-parametric approach. We apply randomization inference—also known as a permutation test—to the hypothesis that the errors in the data from our various sources are *not* category-dependent. Specifically, in each of 999 (unique) permutations, we randomly assign the ground truth category for each task (while preserving the size of each category) and then compute each worker's frequency of correctness conditioned on the (assigned) ground truth category. The test statistic is the median absolute difference between these frequencies. To obtain an exact p-value for this hypothesis test, we compute the number out of all 1000 computed medians[3] that are at least as extreme as the median observed under real categories.

Table 1: Summary of Each Data Set.

	Workers	Tasks	Responses	
			$\text{gt} = 0$	$\text{gt} = 1$
BM	83	1000	2545	2455
HCB	722	3267	8767	10932
RTE	164	800	4000	4000
TEMP	76	462	2590	2030
WB	39	108	2340	1872
WVSCM	17	159	1219	731
SP	143	500	4900	5100

Using this test, we find very strong evidence to *reject* the null hypothesis that errors in each data set are not category-dependent. For nearly every data, the median observed under the real categories is the most extreme, corresponding to a p-value of 0.001 for our test. In the remaining datasets, TEMP and SP, the observed median is still quite extreme, corresponding to p-values of 0.020 and 0.006, respectively.[4] As a result, it is apparent that, even for this diverse collection of tasks with binary categories, category matters a great deal in determining the pattern of errors in crowdsourcing data.

In addition to being statistically significant, the dependence on categories is also *practically* significant. In Table 2, we show the median absolute difference (MAD) between frequencies of correctness conditioned on the ground truth categories for each data set and estimate a 95% confidence interval for these values via bootstrap resampling. We emphasize that these values are *absolute* differences. Large values do not necessarily imply that one category is substantially easier than the other; workers can differ in the category for which their responses are more accurate. Then, we compare the true median absolute differences—our test statistic (TS)—to the median and maximum values of the test statistics observed in the permuted data during our randomization inference. In this comparison, the true value of the test statistic, and even the lower bound of the confidence interval, is often much greater than the maximum value of the test statistic observed in any permutation.

[2] It also lists a seventh data set, SpamCF, which appears to meet our criteria, but upon closer inspection only contains ground truth categories for tasks where the workers were in unanimous agreement.

[3] 999 under permutations of the ground truth categories and 1 under the real ground truth categories from the data.

[4] Further, if the mean absolute difference in frequency of correctness is used in place of the median as the test statistic, then the observed mean is the most extreme value (and, thus, $p = 0.001$) for every data set.

Table 2: Summary of Randomization Inference Results: Testing Null Hypothesis of Category Independence.

	MAD	95% CI	Med TS	Max TS	p
BM	0.382	(0.318, 0.476)	0.100	0.167	0.001
HCB	0.364	(0.333, 0.471)	0.166	0.197	0.001
RTE	0.138	(0.131, 0.200)	0.088	0.111	0.001
TEMP	0.085	(0.050, 0.167)	0.061	0.119	0.020
WB	0.408	(0.319, 0.550)	0.058	0.179	0.001
WVSCM	0.238	(0.148, 0.436)	0.079	0.179	0.001
SP	0.065	(0.053, 0.091)	0.049	0.071	0.006

5 TASK & WORKER HETEROGENEITY

In this section, and further in the supplementary material, we explore the degree to which tasks and workers exhibit heterogeneity with a variety of approaches. For tasks, we say they are heterogeneous if the probability of a correct response from a worker tends to vary with the underlying task (and homogeneous otherwise). For workers, we say they are heterogeneous if the probability of a correct response on a task tends to vary with the worker who is providing the response (and homogeneous otherwise).

5.1 MODEL-AGNOSTIC ANALYSIS

Heterogeneity in Tasks. Previously, we found that there is strong evidence that tasks are heterogeneous based on their category. The next question we consider is whether tasks are homogeneous—i.e., whether workers have a constant probability of correctness—within each category. We once again employ randomization inference to test the null hypotheses that tasks within each ground truth category are homogeneous. Consequently, we perform two hypothesis tests in each data set—one per category. For these tests, our test statistic is the difference in the mean (DiM) frequency of correct responses between apparently difficult tasks and apparently easy tasks in the given category. The apparently difficult and apparently easy tasks are the upper and lower half, respectively, of the set of all tasks in that category when sorted in order fraction of correct responses for each task. We perform the randomization inference by (uniquely) permuting the identifiers of the tasks within the given category (999 times); this preserves the number of times each task appears in the set of all responses, but changes which workers are associated with which tasks. We obtain exact p-values as above. The results of these tests are displayed in Table 3.

For most categories, in most data sets, this test suggests rejecting the null hypothesis of homogeneity. However, in contrast to our randomization inference for categories, there is some reason to be skeptical of the practical significance of some of the results, even when they appear statistically significant. The values of the test statistics for the permutations are surprisingly consistent, even to the point of being nearly invariant for category 0 in the BM data set. As a result, there are certain values for which the difference between the true value of the test statistic in the real data and the values observed in the permutations (as summarized by the median and maximum values in Table 3) are quite small, even though the value in the real data is the most extreme value (and thus the associated p-value is small). For example, the true value in category 1 for both the HCB and RTE data sets is less than 0.06 more than the median of the values from the permutations. This suggests that, although we may reject the null hypothesis of homogeneity, the actual difference between homogeneity and the particular kind of heterogeneity that appears to be present in those data sets may not be very meaningful. We will return to this point in Section 5.2.1, when we test the fit of models with different assumptions about task heterogeneity.

Table 3: Summary of Randomization Inference Results: Testing Null Hypothesis of Task Homogeneity.

	gt	DiM	Med TS	Max TS	p
BM	0	0.235	0.235	0.235	0.998
	1	0.603	0.346	0.386	0.001
HCB	0	0.437	0.354	0.376	0.001
	1	0.338	0.292	0.309	0.001
RTE	0	0.242	0.239	0.265	0.412
	1	0.244	0.193	0.217	0.001
TEMP	0	0.142	0.227	0.262	1.00
	1	0.186	0.194	0.228	0.723
WB	0	0.176	0.111	0.150	0.001
	1	0.266	0.125	0.174	0.001
WVSCM	0	0.351	0.228	0.281	0.001
	1	0.419	0.208	0.281	0.001
SP	0	0.184	0.102	0.119	0.001
	1	0.210	0.109	0.125	0.001

5.2 MODEL-INFORMED ANALYSIS

To further investigate task and worker heterogeneity, we need to move beyond our model agnosticism. Without a model, it is not possible to distinguish between, for example, a group of expert workers who completed a standard set of tasks and a group of average workers who completed a

set of particularly easy tasks. In contrast to prior modeling work, though, we employ models as a means to an end—to answer further questions about the data itself. As a result, we employ the standard version of each model. This allows us to perform exact inferences and to make minimal assumptions, while still capturing the essential features of the model.

5.2.1 Finding the Best Fit

We seek to identify the standard DS or IRT model that provides the best fit to each data set. Then, we can use the estimated parameters of those models to answer deeper questions about the data. In light of our results from Section 4, though, IRT models have an obvious shortcoming—they generally do not incorporate category-dependent errors, except in the special case where the same category is more difficult to label than the other for every worker. To address this shortcoming, we also consider an extension of IRT: category-dependent IRT (CIRT), which is similar to the model proposed by Khattak et al. [2016]. For CIRT, we split each data set into two parts by conditioning on the ground truth and fit the standard IRT models to each part independently.

Estimating DS Parameters. Because all of our data sets contain ground truth categories, fitting the DS model is quite straightforward. We use the maximum likelihood estimate (MLE) given in the original paper by Dawid and Skene [1979]. For each worker, their confusion matrix is completely determined given the diagonal entries, which represent the conditional probabilities of answering correctly given each ground truth category. The estimate for each of these entries is simply that worker's empirical frequency of correctness on the tasks they completed in that category. For practical purposes, we must augment these estimates in two ways. First, if a worker did not complete any tasks in a particular category, we use the population-level frequency of correctness for that category as the estimate. Second, to avoid undefined quantities in our model comparison techniques, we hedge extreme estimates $\hat{p} = 0$ or $\hat{p} = 1$ in the following manner:

$$\hat{p}_h = \frac{1}{2n} + \frac{(n-1)}{n}\hat{p},$$

where n is the number of tasks (and h stands for *hedged*).

Estimating IRT Parameters. Our model fitting techniques for IRT models similarly take advantage of the ground truth labels. Unlike in label aggregation generally, this is the standard setting for IRT—when you are grading a test, you generally need to know the answers. The standard algorithm for fitting an IRT model is to use a *marginal maximum likelihood* (MML) approach [Embretson and Reise, 2000, Sanchez, 2021]. In this algorithm, the item parameters are estimated first by computing an MLE while marginalizing over a population-level distribution of ability parameters that is estimated from the data using a quadrature method. Then, the ability parameters are estimated using MLE given the item parameter estimates.

A major assumption underlying IRT model-fitting procedures is that the correct dimension for the ability parameters is specified. We assume these parameters are unidimensional. Although tests to indicate whether ability parameters in a given data set are plausibly multidimensional have been proposed, those methods are designed for settings where nearly all participants respond to nearly all items. They do not readily generalize to crowdsourcing settings where each worker tends to only complete a small subset of the tasks.

We also limit ourselves to the 1PL and 2PL models for IRT and CIRT. Fitting the 3PL model is too computationally expensive in our data sets, which are large compared to typical IRT data. Further, a limitation of our model fitting software [Sanchez, 2021] is that it is not possible to specify or learn a constant value (i.e., 0.5) for the guessing parameter c when using the model fitting methods for the 1PL and 2PL. Thus, c is fixed at the default value of 0 for our experiments.

Lastly, for our largest data set, HCB, fitting the 2PL and C2PL models is too computationally expensive for 10-fold cross validation (see below). Thus, for that comparison, we limit ourselves to the 1PL models for IRT and CIRT. However, we do use the 2PL models for our other model comparison, the Bayesian information criterion (see below).

Comparing Models. There is no perfect method to compare fit among models, particularly those belonging to different model families. Thus, we apply two different procedures: 10-fold cross validation (10FL) and the Bayesian information criterion (BIC) [Gelman et al., 2013, Ch. 7].

10FL involves splitting the tasks into 10 equal-sized parts. For each part i, we fit each of the models on the 9 other parts and use the estimated parameters to make predictions about the probability of correctness for each worker-task pair in part i. These individual predictions are scored using the *quadratic scoring rule*. If there is a particular worker-task pair for which there is no data from the other parts on which to estimate parameters for the worker, that pair is skipped. Finally, models are evaluated using the sum total of their scores for all individual predictions in all 10 parts.

BIC is an adjusted log-likelihood (LLH) measure that penalizes the inclusion of additional parameters:

$$\text{BIC} = k\log(n) - 2\,\text{LLH},$$

where k is the number of parameters and n is the size of the data set, i.e., the total number of responses.

The results of these comparisons are summarized in Table 4. We find that our two comparison procedures tend to agree, giving us more confidence that we are selecting the best model. We put slightly more weight on cross-validation

Table 4: Summary of Model Fitting Results: Best-Fitting Model for each Data Set.

	10FL	BIC
BM	C1PL	DS
HCB	DS	DS
RTE	DS	DS
TEMP	DS	DS
WB	C1PL	C1PL
WVSCM	C1PL	C1PL
SP	C1PL	C1PL

than BIC, so for the one data set (BM) where there is disagreement, we select C1PL as the best-fitting model. Our results indicate that the data sets differ in terms of how useful it is to model characteristics of tasks beyond the ground truth category. The DS model providing the best fit indicates that tasks are more or less homogeneous, whereas the C1PL model indicates that there is heterogeneity. Given this understanding, it is noteworthy that the results of our randomization inference from Section 5.1 do a fairly good job of predicting the results of our model fitting. The data sets that we find are best fit by the DS model include TEMP, for which our randomization inference suggested we should not reject the hypothesis of homogeneity. Further, RTE and HCB are also best fit by the DS model. In those data sets, we found evidence of heterogeneity, but also found reason to question the practical significance of that apparent heterogeneity. Lastly, in both tests, the results for the BM data set are somewhat mixed.

Employing models, though, allows us to do more than just corroborate the results of our previous test. It allows us to gain additional perspective on task heterogeneity beyond what was possible with our model-agnostic analysis. In particular, we find that, even when there is evidence that tasks are heterogeneous, the complexity of that heterogeneity appears limited—the additional discrimination parameters in the 2PL and C2PL models do not improve model fit.

5.2.2 Examining Experts

We can also use our parameter estimates from the best-fitting models to investigate heterogeneity among workers. A convenient one-dimensional summary of a worker's proficiency is their *logit-probability of correctness*, which can be estimated from the parameters of the best-fitting model. For the DS model, probability of correctness is estimated as the sum over all categories of the product of the estimated probability of correctness for that category and its empirical frequency in the set of tasks. For the C1PL model, probability of correctness is estimated using a Monte Carlo method. First, for each ground truth category of tasks, we compute a kernel density estimate (KDE) for the distribution of difficulties. Then, 500 total samples are drawn from these distributions[5], in proportion to the empirical frequency of the ground truth categories. For each sample, we use the IRT eq. (1) to estimate a probability of correctness. Then, we average the probability of correctness over all 500 samples. Lastly, applying the logit function to the estimated probabilities of correctness is a convenient transformation, because it extends the range of the values from $[0, 1]$ to $(-\infty, \infty)$.

Kernel density estimates (KDEs) of the distributions of logit-probability of correctness, where bandwidths are selected using Silverman's rule [Silverman, 1986, scipy.stats.gaussian_kde, Accessed: Oct. 2022], are displayed in Figure 1. For the data sets best fit by the DS model—HCB, RTE, TEMP—we remove outliers at the extreme values. The extreme values for the most part represent workers who responded correctly to every task they completed.[6] We are comfortable removing these workers as outliers, because their extreme estimated logit-probabilities of correctness are very likely to be an illusion of chance and sparse data. We can substantiate this intuition with the following resampling procedure: Fit a normal distribution to the logit-probabilities excluding extreme values (i.e., the max values in all three data sets and the min value in HCB). For each worker in the data set, draw a logit-probability of correctness from the fitted normal distribution. Then, sample a number of correct responses from a binomial distribution with the corresponding probability of correctness where the number of trials is equal to the number of tasks that worker completed in the data. Using this procedure, it is common to observe that both the number of extreme values and the average number of correct responses from the corresponding workers is greater than in the real data.

These estimated distributions offer insights into the validity of a key assumption of many crowdsourcing papers—that there are *expert* workers. Who should count as an expert is a somewhat ill-defined concept in the literature. Sometimes, experts are a distinct group of participants apart from crowdsourcing workers, who are thought or known to be more reliable. We are more interested in experts within the crowd. But there are still questions about who, if anyone, should be counted as an expert. Is it any worker of above-average proficiency? Or is there something more distinct about an expert?

The first thing to note is that nearly all of the distributions in Figure 1 appear to be somewhat left-skewed. In such distributions, considering an above-average worker to be an expert seems inappropriate—the modal worker is above average. Further, the BM, SP, and WVSCM data sets each appear to have one prominent mode, after which the densities drop off relatively steeply. Thus, even if we were to set some threshold to the right of the mode and to consider any workers beyond the threshold to be experts, the density is

[5] We reuse the same 500 samples for each worker.

[6] For HCB, there are also workers who responded incorrectly on each of their tasks, whom we remove for analogous reasons.

small enough that expertise would appear to be relatively insignificant in these data sets. WVSCM is a possible exception. Its shape is very similar to SP, but it is centered in a region where the inverse logit function changes much more quickly. Thus, the relative difference in probability of correctness between a modal worker and a worker in the right tail in WVSCM is greater than that in SP, which may justify considering expertise as more significant in the WVSCM data.

The remaining data sets are all at least plausibly multimodal. This visual intuition is corroborated with statistical hypothesis tests for unimodality in the supplementary material. Like the unimodal distributions, the plausibly multimodal distributions are mostly left-skewed, with the right-most apparent mode being the largest. The distributions drop off less steeply—they are more dispersed than the unimodal distributions. However, these larger tails are mostly in regions where the inverse logit function changes less quickly, so the change in probability of correctness that occurs in the larger tails is less significant. The WB distribution is the exception to these trends. There, the left-most apparent mode is the largest. Moreover, the distribution is centered in a region where the inverse logit function changes quickly. Thus, the right-most apparent mode can be considered a significant cluster of expert workers, distinct from the cluster of workers near the larger mode. We call this phenomenon *strong expertise* to distinguish it from the weaker notion of experts who are in the upper tail of the largest (apparent) mode.

6 DISCUSSION

In Table 5, we summarize our results intuitively in terms of the strength of evidence for various assumptions we find in each data set.[7] Below, we discuss key implications of those pieces of evidence and their significance for future work:

Workers make errors that are category-dependent. In notable theoretical work [Karger et al., 2011, Liu et al., 2012, Ok et al., 2016], it is assumed that workers have a constant probability of correctness. Our results present a challenge to extend theoretical results beyond this simple setting. If provable guarantees are important for designing better algorithms, then those guarantees should be proven under more realistic assumptions. Our results also indicate that, when invoking the IRT model family, it is wise to adopt a CIRT-style model that allows for category-dependent errors, e.g., as is done by Khattak et al. [2016].

Tasks with the same ground truth category may or may not be heterogeneous. When they are heterogeneous, that heterogeneity appears to have limited complexity. Some data sets were best fit by a CIRT model, others by DS. But when CIRT provided the best fit, it was always the least complex model—C1PL. This suggests that when there is heterogeneity within categories of tasks—i.e., when workers do not have a constant probability of correctness per category—the differences within categories can be represented simply.

Workers appear heterogeneous, with distributions of proficiency that are generally well-characterized by the modal workers. Exceptionally reliable "expert" workers do not appear to play a significant role. In the supplementary material, our model-agnostic analysis finds evidence of worker heterogeneity in one (moderate evidence of heterogeneity) or both (strong evidence) categories of each data set. In our model-informed analysis, where we consider the distributions of logit-probability of correctness, workers in each data set exhibit clear heterogeneity. However, many of the distributions have densities that drop off relatively quickly from the largest mode, suggesting that even the most reliable workers do not report correctly with much higher probability than a relatively typical worker.

No set of assumptions universally characterizes the data sets that we consider. As a result, hierarchical (Bayesian) models like those of Lakkaraju et al. [2015] and Paun et al. [2018], which have hyperparameters to capture the degree of diversity across tasks and workers, are likely to be useful. These models can learn whether workers or tasks are completely homogeneous, completely heterogeneous, or something in between. For example, partitioning workers or tasks into a small set of homogeneous clusters may effectively capture the diversity among them. Hierarchical models infer these kinds of relationships directly from the data when estimating model parameters. Further, we note that our results offer some guidance in applying the hierarchical approach. For example, our results suggest that it would be reasonable to adopt a Gaussian prior for a logit-probability of correctness parameter (or, equivalently a logit-normal prior for a probability-of-correctness parameter), as long as category-dependent errors are properly incorporated.

Moreover, the diversity among data sets that we uncover suggests that the degree to which tasks and workers are heterogeneous is something that should be tested rather than assumed when working in a new domain. Understanding the amount (and form) of heterogeneity has important implications for designing or selecting an aggregation algorithm—since candidate algorithms should be tested on a group of representative data sets—and for subsequently quantifying uncertainty in aggregated labels. A concrete next step for future work is to test state-of-the-art label aggregation algorithms on groups of test data sets—including, but not necessarily limited to, those that we consider—that have apparently similar characteristics according to the evidence we summarize in Table 5 and document the extent to which relative algorithmic performance varies among the groups.

[7] Table 5 gives an intuitive summary of our results; the precise meanings of the terms we use in it are discussed in the supplementary material.

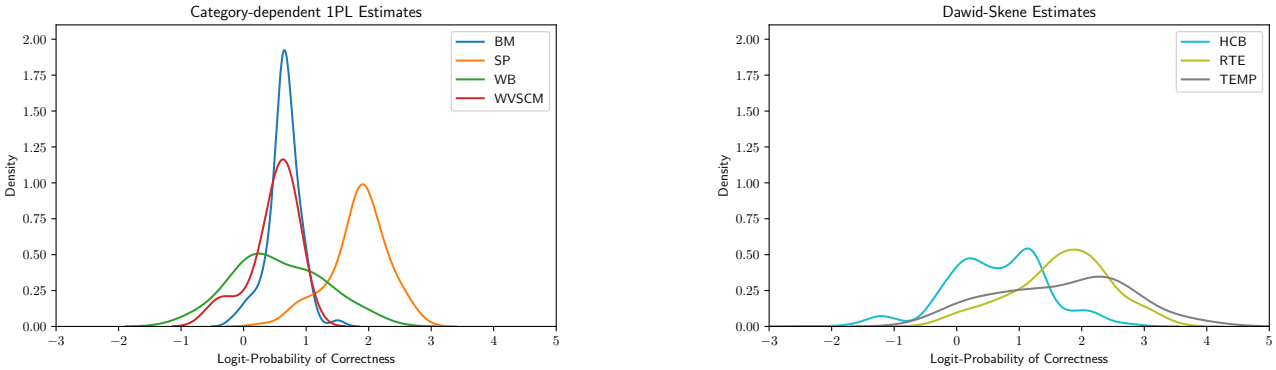

Figure 1: Kernel Density Estimates of the Distributions of Logit-Probability of Correctness in each Data Set.

Table 5: Characterization of Data Sets Based on Strength of Evidence for Assumptions in Experimental Results.

	Category-Dependent Errors	Task Heterogeneity (Intra-Category)	Worker Heterogeneity		Expertise
			Model-Agnostic	Model-Informed	
BM	Very Strong	Moderate	Moderate	Moderate	Weak
HCB	Very Strong	Moderate	Moderate	Strong	Moderate
RTE	Very Strong	Weak	Strong	Moderate	Weak
TEMP	Strong	Weak	Moderate	Moderate	Weak
WB	Very Strong	Strong	Strong	Moderate	Strong
WVSCM	Very Strong	Strong	Strong	Moderate	Weak
SP	Strong	Strong	Strong	Moderate	Weak

Acknowledgements

This work is supported by the National Science Foundation under award #2007256.

References

Yoram Bachrach, Thore Graepel, Tom Minka, and John Guiver. How to grade a test without knowing the answers — a bayesian graphical model for adaptive crowdsourcing and aptitude testing. In *29th International Conference on Machine Learning*, ICML 2012, 2012. arXiv:1206.6386.

Valerio Basile, Federico Cabitza, Andrea Campagner, and Michael Fell. Toward a perspectivist turn in ground truthing for predictive computing, 2021. arXiv:2109.04270.

Alex Burnap, Yi Ren, Richard Gerth, Giannis Papazoglou, Richard Gonzalez, and Panos Y. Papalambros. When crowdsourcing fails: A study of expertise on crowdsourced design evaluation. *Journal of Mechanical Design*, 137(3), 03 2015. ISSN 1050-0472. doi: 10.1115/1.4029065.

Jesus Cerquides, Mehmet Oğuz Mülâyim, Jerónimo Hernández-González, Amudha Ravi Shankar, and Jose Luis Fernandez-Marquez. A conceptual probabilistic framework for annotation aggregation of citizen science data. *Mathematics*, 9(8), 2021. ISSN 2227-7390. doi: 10.3390/math9080875.

A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28, 1979. ISSN 00359254, 14679876.

Susan E. Embretson and Steven P. Reise. *Item Response Theory for Psychologists*. Multivariate Applications Book Series. Psychology Press, 2000. ISBN 9780805828184.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC Texts in Statistical Science. Chapman and Hall/CRC, an imprint of Taylor and Francis, Boca Raton, FL, 3rd edition, 2013. ISBN 9780429113079. doi: 10.1201/b16018.

Mitchell L. Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S. Bernstein. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445423.

Hyun Joon Jung and Matthew Lease. Improving quality of crowdsourced labels via probabilistic matrix factorization. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012. URL https://www.aaai.org/ocs/index.php/WS/AAAIW12/paper/viewFile/5258/5609.

David Karger, Sewoong Oh, and Devavrat Shah. Iterative learning for reliable crowdsourcing systems. *Advances in neural information processing systems*, 24, 2011. NeurIPS'11:4396.

Kohta Katsuno, Masaki Matsubara, Chiemi Watanabe, and Atsuyuki Morishima. Improving reproducibility of crowdsourcing experiments. (Presented in the Work in Progress and Demo track, HCOMP 2019), 2019. URL https://www.humancomputation.com/2019/assets/papers/119.pdf.

Faiza Khattak, Ansaf Salleb, and Anita Raja. Accurate crowd-labeling using item response theory, 03 2016. URL https://www.researchgate.net/publication/299389507_Accurate_Crowd-labeling_using_Item_Response_Theory.

Himabindu Lakkaraju, Jure Leskovec, Jon Kleinberg, and Sendhil Mullainathan. A bayesian framework for modeling human evaluations. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 181–189. SIAM, 2015. doi: 10.1137/1.9781611974010.21.

John P. Lalor, Hao Wu, and Hong Yu. Learning latent parameters without human response patterns: Item response theory with artificial crowds. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, EMNLP-IJCNLP '19, pages 4249–4259, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1434. arXiv:1908.11421.

Yuan Li, Benjamin I. P. Rubinstein, and Trevor Cohn. Truth inference at scale: A bayesian model for adjudicating highly redundant crowd annotations. In *The World Wide Web Conference*, WWW '19, page 1028–1038, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366748. doi: 10.1145/3308558.3313459.

Chao Liu and Yi-Min Wang. Truelabel + confusions: A spectrum of probabilistic models in analyzing multiple ratings. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pages 17–24, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851. arXiv:1206.4606.

Qiang Liu, Jian Peng, and Alexander Ihler. Variational inference for crowdsourcing. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NeurIPS'12, pages 692–700, Red Hook, NY, USA, 2012. Curran Associates Inc. NeurIPS'12:4627.

Barzan Mozafari, Purnamrita Sarkar, Michael J. Franklin, Michael I. Jordan, and Samuel Madden. BM data set. URL https://github.com/ipeirotis/Get-Another-Label/tree/master/data/BarzanMozafari.

Barzan Mozafari, Purnamrita Sarkar, Michael J. Franklin, Michael I. Jordan, and Samuel Madden. Active learning for crowd-sourced databases, 2012. arXiv:1209.3686.

Barzan Mozafari, Purnamrita Sarkar, Michael J. Franklin, Michael I. Jordan, and Samuel Madden. Scaling up crowd-sourcing to very large datasets: A case for active learning. *PVLDB*, 8(2):125–136, 2014.

Jungseul Ok, Sewoong Oh, Jinwoo Shin, and Yung Yi. Optimality of belief propagation for crowdsourced classification. In *International Conference on Machine Learning*, pages 535–544. PMLR, 2016. arXiv:1602.03619.

Naoki Otani, Toshiaki Nakazawa, Daisuke Kawahara, and Sadao Kurohashi. IRT-based aggregation model of crowdsourced pairwise comparison for evaluating machine translations. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 511–520, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1049.

Rebecca J. Passonneau and Bob Carpenter. The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics*, 2:311–326, 10 2014. ISSN 2307-387X. doi: 10.1162/tacl_a_00185.

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. Comparing Bayesian Models of Annotation. *Transactions of the Association for Computational Linguistics*, 6:571–585, 12 2018. ISSN 2307-387X. doi: 10.1162/tacl_a_00040.

Barbara Plank. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.731.

Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *J. Mach. Learn. Res.*, 11: 1297–1322, aug 2010. ISSN 1532-4435. URL http://jmlr.org/papers/v11/raykar10a.html.

Mark D. Reckase. *Multidimensional Item Response Theory*. Statistics for Social and Behavioral Sciences. Springer, New York, NY, 1 edition, 2009. ISBN 978-0-387-89975-6. doi: 10.1007/978-0-387-89976-3.

Ryan Sanchez. Girth: G. item response theory, November 2021. URL https://github.com/eribean/girth. Version 0.8.0.

Documentation: scipy.stats.gaussian_kde, Accessed: Oct. 2022. URL https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.gaussian_kde.html.

Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 614–622, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605581934. doi: 10.1145/1401890.1401965.

Aashish Sheshadri and Matthew Lease. Square: A benchmark for research on computing crowd consensus. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 1(1):156–164, Nov. 2013. URL https://ojs.aaai.org/index.php/HCOMP/article/view/13088.

Bernard W Silverman. *Density Estimation for Statistics and Data Analysis*, volume 26 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, London, 1986.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. RTE and TEMP data sets. URL https://github.com/TrentoCrowdAI/crowdsourced-datasets/tree/master/binary-classification.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 254–263, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

SQUARE. Links to data sets, Accessed: Oct. 2022. URL https://ir.ischool.utexas.edu/square/data.html.

James Surowiecki. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. Doubleday, 1st ed. edition, 2004.

Matteo Venanzi, William Teacy, Alexander Rogers, and Nicholas Jennings. Sentiment popularity data set, 2015. doi:10.5258/SOTON/376544.

Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=TBWA6PLJZQm.

Peter Welinder and Pietro Perona. Online crowdsourcing: Rating annotators and obtaining cost-effective labels. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 25–32, June 2010. doi: 10.1109/CVPRW.2010.5543189.

Peter Welinder, Steve Branson, Serge Belongie, and Pietro Perona. WB data set. URL https://github.com/welinder/cubam/tree/public/demo/bluebirds.

Peter Welinder, Steve Branson, Pietro Perona, and Serge Belongie. The multidimensional wisdom of crowds. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23, pages 2424–2432. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper/2010/file/0f9cafd014db7a619ddb4276af0d692c-Paper.pdf.

Jacob Whitehill, Paul Ruvolo, Tingfan Wu, Jacob Bergsma, and Javier Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, NeurIPS'09, page 2035–2043, Red Hook, NY, USA, 2009. Curran Associates Inc. ISBN 9781615679119. URL https://proceedings.nips.cc/paper/2009/file/f899139df5e1059396431415e770c6dd-Paper.pdf.

Ming Yin, Mary L. Gray, Siddharth Suri, and Jennifer Wortman Vaughan. The communication network within the crowd. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 1293–1303, Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee. ISBN 9781450341431. doi: 10.1145/2872427.2883036.

Yuchen Zhang, Xi Chen, Dengyong Zhou, and Michael I Jordan. Spectral methods meet em: A provably optimal algorithm for crowdsourcing. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. arXiv:1406.3824.

Dengyong Zhou, Sumit Basu, Yi Mao, and John Platt. Learning from the wisdom of crowds by minimax entropy. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper/2012/file/46489c17893dfdcf028883202cefd6d1-Paper.pdf.

Dengyong Zhou, Qiang Liu, John C. Platt, Christopher Meek, and Nihar B. Shah. Regularized minimax conditional entropy for crowdsourcing, 2015. arXiv:1503.07240.