Testing Conventional Wisdom (of the Crowd) Supplementary Material

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

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

1 HETEROGENEITY OF WORKERS: MODEL-AGNOSTIC ANALYSIS

Perhaps the most ubiquitous assumption in label aggregation—which is rarely even acknowledged as an assumption—is that workers vary in their proficiency, e.g., by having different probabilities of correctness than other workers (when completing a task in a given category). This assumption is not universal, however. For example, the image classification error models discussed by Wei et al. [2022] assume that workers are homogeneous.

To explore this assumption, we once again employ randomization inference. This time, we test the null hypothesis that workers are homogeneous when completing tasks in the same category. The test is very similar to our randomization inference concerning heterogeneity of tasks in the main paper. We perform two hypothesis tests in each data set—one for each category. For these tests, our test statistic is the difference in the average frequency of correct responses between apparently more proficient workers and apparently less proficient workers in the given category. The apparently more proficient workers and less proficient workers are the upper and lower half, respectively, of the set of all workers when sorted in order of each worker's fraction of correct responses in the given category. To perform the randomization inference, we (uniquely) permute the identifiers of the workers within the given category, thereby preserving the number of times each worker appears in the set of all responses, but changing which tasks are associated with which workers (999 times). Lastly, to obtain an exact p-value, as in the main paper, we calculate the number of test statistics out of 1000 that are at least as extreme as the true test statistic from the original data. The results of these tests are displayed in Table 1.

We find that for each data set, in at least one category, there is strong evidence to reject the null hypothesis that workers are homogeneous. Unexpectedly, there are some data sets (BM, HCB, and TEMP) for which this does not hold for both categories. However, there are reasons to interpret this result cautiously. As always, lack of evidence against the null hypothesis does not necessarily constitute evidence for it; in this case, we believe the null hypothesis is *a priori* unlikely. A possible explanation for these results that would not necessarily support the null hypothesis would be that for some categories, in some data sets, workers did not complete enough tasks for it to be clear that they have heterogeneous proficiency.

Another way in which the results from this test are somewhat weaker than those from our randomization inference about task heterogeneity is that they are not obviously corroborated by our model-informed analysis. In the plot of logit-probabilities of correctness in the main paper, the BM data set does appear to have the least dispersion among the distributions of logit-probability of correctness, but it is also dense in a region where probability of correctness changes quickly with changes in logit-probability of correctness. The HCB and TEMP data sets, on the other hand, have relatively high dispersion. This does not necessarily contradict the results of our tests—logit-probability of correctness incorporates worker proficiency in both categories, whereas the randomization inference indicates a lack of support for heterogeneity in just one category in each data set. However, it does not clearly corroborate the results either.

2 A NOTE ON DIMENSIONALITY OF IRT ABILITY PARAMETERS

As is discussed in the main paper, a major assumption underlying IRT model-fitting procedures is that the correct dimension for the ability parameters is specified. The IRT literature includes a few procedures that are designed to indicate whether

Table 1: Summary of Randomization Inference Results: Testing Null Hypothesis of Worker Homogeneity.

	gt	DiM	Med TS	Max TS	p
BM	0	0.226	0.188	0.316	0.103
	1	0.469	0.384	0.456	0.001
HCB	0	0.649	0.534	0.564	0.001
	1	0.391	0.430	0.468	0.999
RTE	0	0.273	0.215	0.261	0.001
	1	0.229	0.183	0.220	0.001
TEMP	0	0.254	0.237	0.307	0.197
	1	0.359	0.236	0.304	0.001
WB	0	0.381	0.089	0.130	0.001
	1	0.462	0.113	0.164	0.001
WVSCM	0	0.398	0.175	0.282	0.001
	1	0.318	0.183	0.297	0.001
SP	0	0.178	0.134	0.204	0.009
	1	0.188	0.138	0.201	0.002

Table 2: Summary of Modality Test Results: Null Hypothesis of Unimodality.

	Dip Test	BW Test
BM	0.698	0.110
HCB	<**0.001**	**0.037**
RTE	<**0.001**	0.324
TEMP	**0.028**	0.379
WB	**0.011**	0.224
WVSCM	0.970	0.192
SP	0.259	0.616

ability parameters in a given data set are plausibly multidimensional, but those methods are designed for settings where nearly every participant responds to nearly every item. They do not readily generalize to crowdsourcing settings where each worker tends to only complete a small subset of the tasks. We attempted to adapt one such procedure—DIMTEST (See [Reckase, 2009, Ch. 7])—to our setting, but the resulting procedure failed to reliably distinguish between synthetic data generated using unidimensional and multidimensional IRT models.

3 MODALITY TESTING THE EMPIRICAL DISTRIBUTIONS OF LOGIT-PROBABILITY OF CORRECTNESS

Statistical hypothesis tests for unimodality of the empirical distributions of logit-probabilities (not of the KDEs for those distributions that we plot in the main paper)—calibrated versions of Hartigan's dip test and Silverman's bandwidth test [Johnsson et al., 2017, 2018]—confirm the visual intuition from our analysis in the main paper that certain distributions of logit-probabilities of correctness are plausibly multimodal. The results of these statistical tests are presented in Table 2. Specifically, plausible multimodality (i.e., the rejection of the null hypothesis of unimodality) under these tests indicates that the smaller apparent modes would be unlikely to result from random chance if the true underlying distributions were unimodal.

4 HETEROGENEITY OF WORKERS: MODEL-INFORMED ANALYSIS

Multi-modality (or plausible multi-modality) in the distribution of logit-probability of correctness suggests that workers are heterogeneous, i.e., they have different probabilities of correctness. In testing the null hypothesis of unimodality for distributions of logit-probability of correctness (Section 3), however, there were three data sets (BM, WVSCM, and SP) for which the evidence did not suggest that we should reject the null hypothesis of unimodality. Those three data sets were all fit best by the C1PL model. So, for those distributions, we use the C1PL model to construct a model-informed test of the null

Table 3: Summary of Model-Informed Resampling Test Results: Null Hypothesis of Worker Homogeneity.

	Observed Variance	Med TS	Max TS	p
BM	0.065	0.035	0.069	0.001
WB	0.481	0.030	0.057	0.001
WVSCM	0.135	0.037	0.128	0.001
SP	0.220	0.095	0.154	0.001

hypothesis of heterogeneity that does not involve modality.

The test is a model-informed resampling procedure. First, we estimate the parameters of the C1PL model using marginal maximum likelihood estimation (as in the main paper). Then, we resample each worker's responses to each task that they responded to in the real data set according to the estimated C1PL model (999 times). The parameters for each task in that model are assumed to be those that were estimated from the data. The ability parameters in each category for each worker in that model are assumed to be equal to the *average* of the ability parameters in that category that were estimated from the data. Thus, workers are assumed to be homogeneous.

Using the simulated data from each round of resampling, we estimate the empirical distribution of logit-probability of correctness as in the main paper. For our test statistic, we use the variance of the distribution of logit-probability of correctness. Thus, we compare the variance of the simulated distributions to the value for the variance that we observe in the real data.

Results are presented in Table 3. In all three data sets, the observed variance is more extreme than the variance of any distribution resulting from simulation under the null hypothesis of homogeneity ($p = 0.001$). Thus, the results of this test provide evidence against that null hypothesis. (Additionally, the result is the same for the WB data set, which was also fit best by the C1PL model, but was found to be plausibly multimodal in Section 3, above.)

5 FURTHER EXPLORING TASK HETEROGENEITY: DIABOLICAL TASKS

In a setting with strong expertise (see the main paper), a natural question arises. How much are experts worth relative to a regular worker? In many cases, if it is costly to recruit or identify experts, then doing so might not be worth it. Aggregating the responses from a few non-expert workers may be cheaper and just as, if not more, accurate. However, it is not difficult to imagine cases where experts provide additional value. For example, they may have domain-specific knowledge that non-experts do not possess that leads them to produce correct responses even when the majority of non-experts fails to do so. That is, there may be cases where the aggregation of non-experts will fail to identify the correct category, but an expert will succeed. More generally, we refer to the kind of task where non-experts tend to respond incorrectly, but experts tend to respond correctly, as a *diabolical task*.

We search for possible diabolical tasks in the WB data set. First, we fit a Gaussian Mixture Model (GMM) [sklearn.mixture.GaussianMixture, Accessed: Oct. 2022] to the logit-probabilities of correctness that we computed in the main paper in order to classify workers as either experts or non-experts. Then, we look for tasks that meet the following criteria:

1. At least two experts and non-experts completed the task.
2. A majority of non-experts produced an incorrect response.
3. A majority of experts produced a correct response.

There are 27 tasks that meet these criteria—25% of all tasks. This is a substantial number, but there are a few unusual features of the WB data set that may somewhat temper its significance. Most importantly, the relative frequency of experts is quite high. As a result, it is not uncommon for the majority of all workers to respond correctly even when the majority of non-experts responds incorrectly. This occurs for 17 out of the 27 apparently diabolical tasks. Also, relative to modal workers in the other data sets, the non-expert workers in WB perform fairly poorly. These mitigating factors suggest that the significance of diabolical tasks for label aggregation in this particular data set is likely narrow. More generally, diabolical tasks may be a bigger factor in settings where experts are a population distinct from crowd workers and, thus, may be more likely to differ from crowd workers in systematic ways.

Alongside our analyses of worker heterogeneity and expertise in the main paper, the discussion of diabolical tasks suggests another key implication of our analysis:

Relying on the existence of experts who can be reliable even when the majority is unreliable may be misguided. Overall, we find that it is often the case that the most reliable workers are not much more reliable than a relatively typical (modal) worker. Further, it can be argued in some cases that the improvement in probability of correctness for an “expert” worker does not fully compensate for their decreased frequency in the population. For example, consider a single expert worker, who is more proficient than a modal worker, and whose logit-probability of correctness corresponds to a density that is about one third of the density at the largest (approximate) mode according to the KDE for the distribution of logit-probability of correctness. If such an expert is less likely to produce a correct response than a majority of 3 workers, each with the (approximate) modal logit-probability of correctness, then the additional value provided by the expert worker may not be worth the additional cost of identifying them. Moreover, the modal workers are often both reliable and plentiful, meaning that their responses can be aggregated into very reliable labels. This corroborates the work of Li et al. [2019], who find that majority vote is a powerful aggregation algorithm on real crowdsourcing data.

6 DISCUSSING TERMS USED IN TABLE 5 OF THE MAIN PAPER

Category-Dependent Errors. **Strong** means that the p-value for using randomization inference to test the null hypothesis of category-independent errors was below 0.05. **Very Strong** means that the observed test statistic was more extreme than every test statistic generated under the randomization inference procedure.

Task Heterogeneity (Intra-Category). **Weak** means that the p-value for using randomization inference to test the null hypothesis of task homogeneity was above 0.05 in at least one category and the data were best fit by the DS model according to both fit comparisons (10FL and BIC). **Moderate** means that either the p-value for using randomization inference to test the null hypothesis of task homogeneity was below 0.05 in both categories, despite the DS model providing the best fit for the data (as in the case of HCB) or that the p-value for using randomization inference to test the null hypothesis of task homogeneity was below 0.05 in at least one category and the data were best fit by a CIRT model according to at least one fit comparison[1] (as in the case of BM). **Strong** means that the observed test statistic was more extreme than every test statistic generated under the randomization inference procedure and that the data were best fit by a CIRT model according to both comparisons.

Worker Heterogeneity, Model-Agnostic. **Moderate** means that the p-value for using randomization inference to test the null hypothesis of worker homogeneity was below 0.05 in at least one category. **Strong** means that the p-value for using randomization inference to test the null hypothesis of worker homogeneity was below 0.05 in both categories.

Worker Heterogeneity, Model-Informed. **Moderate** means either that the p-value for testing the null hypothesis of unimodality of the estimated distribution of logit-probabilities of correctness was below 0.05 for one of the modality tests (as in the case of RTE, TEMP, and WB) or that the p-value for testing the null hypothesis of worker homogeneity using model-informed resampling (Section 4) was below 0.05 (as in the case of BM, WB, WVSCM, and SP). **Strong** means that the p-value for testing the null hypothesis of unimodality of the estimated distribution of logit-probabilities of correctness was below 0.05 for both of the modality tests.

Expertise. **Weak** means that the estimated distributions of logit-probability of correctness were either apparently unimodal or plausibly multimodal (according to one modality test, but not both) with density that drops off relatively quickly from the largest mode, which is also the right-most apparent mode. **Moderate** means that the estimated distributions of logit-probability of correctness were plausibly multimodal according to both modality tests (as in the case of HCB). **Strong** means that the estimated distributions of logit-probability of correctness were plausibly multimodal (according to at least one modality test) with the largest mode not being the right-most apparent mode (as in the case of WB).

7 SOFTWARE

Our code, available at `https://github.com/burrelln/Testing-Conventional-Wisdom`, is implemented in Python 3. To fit IRT models using the standard marginal maximum likelihood (MML) technique, we use the G. Item

[1] Particularly if the method of comparison for which a CIRT model provided the best fit were 10FL, to which we give slightly more weight than BIC.

Response Theory (girth) package [Sanchez, 2021]. To perform calibrated statistical hypothesis tests for the unimodality of empirical distributions, we use the modality package [Johnsson et al., 2018]. In order for this package to work in Python 3, we had to modify the source code. In particular, it was necessary to change the print statements from the Python 2 syntax to the Python 3 syntax.

The rest of our tests and procedures were implemented by us. They rely on the following well-known Python packages: numpy [Harris et al., 2020], pandas [pandas development team, 2020, Wes McKinney, 2010], scikit-learn [Pedregosa et al., 2011], and scipy [Virtanen et al., 2020]. The logit-probabilities of correctness in the main paper were plotted using the seaborn package [Waskom and the seaborn development team, 2020].

References

Charles R. Harris, K. Jarrod Millman, St'efan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fern'andez del R'ıo, Mark Wiebe, Pearu Peterson, Pierre G'erard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2.

Kerstin Johnsson, Magnus Linderoth, and Magnus Fontes. What is a "unimodal" cell population? using statistical tests as criteria for unimodality in automated gating and quality control. *Cytometry Part A*, 91(9):908–916, 2017. doi: 10.1002/cyto.a.23173.

Kerstin Johnsson, Russell Jarvis, Avinash Varna, and Tom Pollard. modality, 2018. URL https://github.com/kjohnsson/modality. Version 1.1.

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.

The pandas development team. pandas-dev/pandas: Pandas, February 2020.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

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: sklearn.mixture.GaussianMixture, Accessed: Oct. 2022. URL https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Michael Waskom and the seaborn development team. mwaskom/seaborn, September 2020. URL https://doi.org/10.5281/zenodo.592845.

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.

Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56–61, 2010. doi: 10.25080/Majora-92bf1922-00a.