# How Many Ratings per Item are Necessary for Reliable Significance Testing?

**Anonymous ACL submission**

## Abstract

Most machine learning evaluation is based on the assumption that machine and human responses are repeatable enough to be measured against data with unitary, authoritative, "gold standard" responses, via simple metrics such as accuracy, precision, and recall. However, AI models have multiple sources of stochasticity. And the human raters who create gold standards tend to disagree with each other, often in meaningful ways. Thus, a single output response per input item may not provide enough information. We introduce methods for determining whether an (existing or planned) evaluation dataset has enough responses per item to reliably compare the performance of one model to another. We apply our methods to several of very few extant gold standard test sets with multiple disaggregated responses per item and show that there are usually not enough responses per item to reliably compare the performance of one model against another. Our methods also allow us to estimate the number of responses per item for hypothetical datasets with similar response distributions to the existing datasets we study. When two models are very far apart in their predictive performance, fewer raters are needed to confidently compare them, as expected. However, as the models draw closer, we find that a larger number of raters than are currently typical in annotation collection are needed to ensure that the power analysis correctly reflects the difference in performance.

## 1 Introduction

A design problem that is critical for assuring end-to-end quality in AI is to determine how much data must be collected to ensure that the breadth of the operational domain is fully tested. This is one of the problems that statistical *power analysis (PA)* solves (Bausell and Li, 2002). After tests are run, one must determine how reliable the results are, based on the amount of data collected. Here, *null hypothesis statistical tests (NHSTs* and *confidence intervals (CIs)* are widely used when reporting experimental results.

NHST and CI, in the context of PA, are widely considered to be gold standards for estimating variance due to sampling error; however, nearly all existing approaches for them fail to capture a key source of variance in AI systems, namely *response variance*. This comes from two sources: models and humans. During AI model inference, non-determinism arises from parallelism (floating point math on different data arrival orders (Shanmugavelu et al., 2024), MoE routing (Shazeer et al., 2017)) in addition to inherent stochasticity in inference process (Monte Carlo dropout (Gal and Ghahramani, 2016), ensembling (Lakshminarayanan et al., 2017)) as well as the model itself (variational autoencoders, *temperature* parameter of LLMs).

Human feedback continues to play a critical role in making AI useful. The increasingly sophisticated behavior of AI models has made it easier for people with little-to-no computer training to interact with them (Daugherty and Wilson, 2018). However, humans themselves frequently disagree in their views and behaviors, and on "gold standard" annotations. This is true on important but highly subjective questions such as what is offensive, but also for much more seemingly objective or mundane tasks such as medical image analysis or object detection.

We present a method, building on the multistage bootstrapping approach of (Wein et al., 2023), for estimating the amount of test items $N$, *and responses per item* $K$, to account for sampling variance in AI evaluation with humans in the loop, *before more data are collected and models are retrained*, thus giving us critical information about how to budget resources for building benchmark datasets. This approach simulates the responses from a pool of human raters and two ML models,

rather than relying on analytical methods that are not designed for multistage sampling. The simulation enables us to generate enough response data to explore the significance boundary for NHST under various metrics, for $N$ test examples (items) with $K$ responses per item for each model and pool of human raters. We apply this method to two existing evaluation datasets that have multiple responses per item, and have retained those individual ratings rather than an aggregate response per item (e.g., per-item means). We show that many of these datasets lack enough responses to construct reliable evaluations of ML models. We further show that, in general, one can create reliable datasets with fewer total responses by collecting *fewer items with more responses per item*.

## 2 Related Work

Statistical testing of system performance is critical to the understanding of state-of-the-art performance on a task or within a domain, in particular due to the flawed nature of benchmarking practices in machine learning evaluation (Ethayarajh and Jurafsky, 2020; Raji et al., 2021; Rodriguez et al., 2021; Hernandez-Orallo, 2020). Existing metrics such as Student's t-test (Student, 1908) are based on the assumption that the datasets are normally distributed with the same standard deviation (Søgaard et al., 2014), which may not be the case machine learning predictions and are therefore not applicable, in particular when testing the system on new datasets (Søgaard, 2013). Dietterich (1998) applied hypothesis testing to machine learning systems and (Dror et al., 2020; Deutsch et al., 2021) provide a survey and guide to state-of-the-art techniques for statistical significance testing in AI systems. Longjohn et al. (2025) study the problem of aggregating across multiple tests.

All of these studies apply to the case where each model yields a single response and a single correct label exists for each training example; therefore, the issue of response variance is ignored. More recently, Gundersen (2020) exploited pseudo-random seeds to generate multiple model responses that could be used for improved statistical testing in the presence of a single correct label for each item. Goldberg et al. (2018) showed how to revise $p$-value calculation when "gold" annotations exist but are unknown and in its place multiple noisy "bronze" annotations are available, where the probability of a bronze annotation matching the gold is given. In our work, ratings are subjective and, hence, there is no single right answer; that is, ground truth is a distribution.

Our approach incorporates response variance from both ML models and human raters. The nature of response variance of the former was studied in (Szymański and Gorman, 2020), claiming that human rater response variance on individual items is most often due to measurable differences in perspective or ambiguity of the item, as opposed to noise. The nature of response variance in ML models was studied in (R Artstein, 2008; Plank et al., 2014).

Related crowdsourcing studies have examined the trade-offs between cost and quality of annotation collection (Snow et al., 2008) or gave recommendations for which crowdsourcing platforms and protocols to use (Wang et al., 2013). Chau et al. (2020) explored the use of peer-review and self-review to resolve disagreement in annotating, and Hovy et al. (2013) developed an unsupervised model to identify which Mechanical Turk raters are reliable. Recent assessments of leaderboard practices have also led to models able to indicate which items are most useful to annotate for evaluation purposes (Rodriguez et al., 2021). (Welinder and Perona, 2010) developed a system to select the most useful/informative labels to collect, which can lead to a reduction in annotation cost.

Sheng et al. (2008) focus on ML data curation and examines when one should obtain multiple, noisy training labels to improve model accuracy, assuming there exists a single correct label for each example. Lin et al. (2014) claims that response variance is less important than item variance – at least for training data – and suggests collecting more items with a single response is more valuable than collecting multiple responses per item.

Wein et al. (2023) investigate $p$-value sensitivity of both metrics and test-set sampling methods in hypothesis testing, which therefore can affect the power analysis. While the latter did not turn out to be important in our study, metrics did. Clearly, different metrics (e.g., mean absolute error vs Spearman rank-correlation) will produce different scores for the same matrix of responses, so it stands to reason that any comparison will have different $p$-values for different metrics. They model a metric as a function $\Gamma(M, G)$, where $M$ is a matrix of model predictions which returns a score for $M$. We assume $\Gamma$ is given here but focus on the best per-

forming of these metrics in experiments. Homan et al. (2024) initiates a study of the trade-off between number of items and responses using a toy simulator. Here, we use real datasets to investigate these trade-offs and perform experiments that shed light on the mechanism for how response variance provides statistical significance.

The term *multistage sampling* is commonly used in statistics when the data is subsampled at multiple levels of granularity, usually for stratification. Bootstrap resampling has been applied in this setting (Mashreghi et al., 2016) and so the sampling method we describe herein can be seen as an instance of these. The Pigeonhole Bootstrap (Owen, 2007) is quite different from our multistage bootstrapping in that it resamples independently over rows and columns to form a Cartesian product rather than being nested.

It would be remiss not to mention other classes of techniques besides hypothesis testing that are commonly used for measuring statistical differences in model performance; see (Riezler and Hagmann, 2021) for a survey. *Likelihood ratios* provide an alternative form of significance testing and have been used for evaluating the impact of variability in data characteristics and hyperparameter settings on ML models (Hagmann et al., 2023). Estimation statistics for reliability, most notably *confidence intervals*, take variance into account to produce a range of values and are often used to assess a difference in model performance via non-overlap. *Circularity testing* based on general additive models has been proposed for evaluating the validity of ML models (Riezler and Hagmann, 2021).

## 3 Problem Statement

We wish to apply null hypothesis significance testing (NHST) to compare the performance of two ML models, $A$ and $B$, on a test set of $N$ items with $K$ responses per item and decide if one model is significantly better than the other. We evaluate this with respect to human-annotated benchmark "gold" responses, $G$, and according to a metric, $\Gamma$, which we assume is provided as a design hyperparameter. For example, a common metric for evaluating regression models is the mean absolute error (differences) between model predictions and gold annotations.

The null hypothesis makes the assumption that the respective model output distributions are the same in relation to $G$. Our goal is to determine whether the observations would be at most $5\%$ likely under the null hypothesis and, therefore, the null hypothesis can be rejected. The $5\%$ level is what our calculated $p$-values are compared against to conclude significance. Our motivation here is to determine whether a dataset—which we represent as $G^{N\times K}$, a matrix of $N$ items and $K$ responses—is large enough to provide replicable test results. This can be applied either post-hoc, as a test of the reliability of results, or at design time, before data is gathered and to help determine how best to allocate the usually limited amount of resources available for gathering human annotations.

A key innovation in this work is to treat a data set $G^{N\times K}$ (as well as the responses from models $A$ and $B$) as a matrix of responses, instead of the pervasive simplifying assumption that $G$ is a vector, whose value for each item is an aggregation, such the mean, of a number of independent annotator (or model) responses. The notation captures the further insight that the distribution of responses for each item in a dataset is different.

## 4 Methods

### 4.1 Simulator

We use a simulator to generate annotations and model predictions for individual items by modeling the responses for each item as a random variable.

It takes input parameters $N$ and $K$, along with a *perturbation parameter* $\epsilon$. In the first stage, it randomly chooses hyperparameters $\theta_1, \ldots, \theta_N \sim P_{items}$, each corresponding to an item $\theta_i$, from a fixed distribution that serve as model parameters for the second stage. In the second stage, for each item $i$ we sample $K$ responses from a second distribution $P_{responses}(\theta_i)$. We do this for each of three datasets respectively representing responses from gold annotations, $G^{N\times K}$ and two machines, $A^{N\times K}$ and $B^{N\times K}$.

These choices operationalize a solution to the paradox that one must have data in $G$, $A$, and $B$ to know if it has enough statistical power. Instead, we simulate a set of gold items and responses ($G$) and then simulate an ideal machine ($A$) – ideal because it draws its simulated responses from the same distribution as the gold – and then explore how such an ideal system would compare in significance to another model ($B$) whose response distributions differ from gold by an amount ($\epsilon$) we experimentally control. This gives us *a-priori* control over the hypothesis test, because we know which model

is better through a controllable parameter.

For any given selection of $N$ and $K$, we have response matrices $G^{N\times K}$ and $A^{N\times K}$ and, for each $\epsilon$, a matrix $B^{N\times K,\epsilon}$. We then seek to compare $A$ and $B$ to each other to determine which is better; the answer, should almost always be $A$ unless $\epsilon = 0$. When evaluating AI systems, the comparison of $A$ and $B$ involves differencing each of their item responses to those of $G$ using a suitable metric, which is then aggregated across the items. We compare the performance between $A$ and $B$ via $\Gamma(A, B, G)$.

### 4.2 Estimating $p$-values

Given $N$, $K$, and $\epsilon$, $p$-values are estimated by drawing $b$ (bootstrap) resamples $S_{alt} = \langle G_1^{N\times K}, A_1^{N\times K}, B_{1,\epsilon}^{N\times K}\rangle, \ldots, \langle G_b^{N\times K}, A_b^{N\times K}, B_{b,\epsilon}^{N\times K}\rangle$ for the alternative hypothesis according to the process described in Section 4.1. Since the null hypothesis makes the assumption that the distributions of $A$ and $B$ are the same with respect to $G$, we construct $S_{null}$ by pooling the items from $A^{N\times K}$ and $B^{N\times K}$ and then independently sampling from this pool. When sampling responses from $A$, for each item $i$ we sample each response by sampling from $P_{responses}(\theta_i)$, where $\theta_i = (\mu_i, \sigma_i)$. Sampling responses from $B$ is similar but we first choose $\delta_i \sim Unif(-\epsilon, \epsilon)$ and then sample from $P_{responses}(\theta_i)$, where $\theta_i = (\mu_i + \delta_i, \sigma_i)$.

Next, we estimate the *expected p-value under the alternative hypothesis* as the average one-sided $p$-value over all samples in $S_{alt}$, computed by counting for each $s_{alt} = \langle G_{alt}^{N\times K}, A_{alt}^{N\times K}, B_{alt,\epsilon}^{N\times K}\rangle \in S_{alt}$ the fraction of samples $s_{null} \in S_{null}$ where $\Gamma(s_{null})$ is at least as extreme as $\Gamma(s_{alt})$. Here "at least as extreme" is determined by computing $\Gamma_{alt}$ (respectively, $\Gamma_{null}$), the median of $\Gamma$ over $S_{alt}$ (respectively, $S_{null}$). If $\Gamma_{alt} > \Gamma_{null}$, then "at least as extreme" means $\Gamma(s_{null}) \geq \Gamma(s_{alt})$. Otherwise, it means $\Gamma(s_{null}) < \Gamma(s_{alt})$. The estimator is fast to compute if the $\Gamma$ values are presorted, and because it is averaged over a large number of samples from the alternative hypothesis, it is a robust estimator for determining whether $N \times K$ is a large enough sample size.

Finally, as is typical for NHST, we reject the null hypothesis when the $p$-value is below significance level $\alpha = 0.05$.

### 4.3 Fitting the Simulator to Real Data

The simulator allows us to generate many test sets to extrapolate patterns beyond one domain or system. By holding the item distributions for $A, B$ and $G$ fixed, we can draw from them repeatedly to generate test sets similar to a real dataset but with arbitrarily large values of $N$ and $K$, which would be infeasible with actual human annotations.

In contrast to the pure simulation framework from (Wein et al., 2023), the datasets we study have discrete-valued responses. Therefore, in order to apply the simulator framework to these datasets, we use per-item *location* and *scale* measures (e.g., mean and standard deviation) to fit distributions — one for location and one for scale — so that we can draw samples $\{(\mu_i, \sigma_i), i \in [1, N]\}$ and then sample an item's responses from a *generalized normal distribution* $\mathcal{N}(\mu_i, \sigma_i)$.[1] While distribution fitting is outside the scope of this paper, one can employ the simple technique of computing per-item means and standard deviations and then using grid search on hyperparameters for $P_{items}$ to minimize the expected mean absolute error between simulated vs real per-item location and scale values.

We used the *censored* normal distribution for $\mathcal{N}$, which assumes a latent continuous distribution that is not observed exactly but measured to within intervals, including left and right intervals which *pool* (not truncate) the smallest and largest values, respectively. This provides support for head and/or tail bias. For example, items in the Stanford Toxicity dataset (see Section 5) rated at either extreme (either "not toxic" or "extremely toxic") tend to have more agreement among raters. We use distributions fitted to each dataset from distribution families tailored to each dataset. This involves visualizing the distributions of response means and standard deviations of the item responses in each dataset to get a sense of what they look like and then choosing a parameterized family of distributions to fit the data to. Figure 1 illustrates goodness-of-fit for simulations of datasets used in this paper.

## 5 Data

Unfortunately, there are precious few public datasets having both a large number of items and disaggregated responses. We apply the metrics and $p$-value estimators to the following datasets, all

[1] This framework is general enough to accommodate an additional *shape* parameter, such as the skew of a skew normal distribution, though it wasn't utilized in our experiments.

of which are secondary to us. We essentially ignore the content of each item in each dataset and use only the human responses associated with each item. Even though these responses were generated by humans–and we believe modeling human annotators is a promising direction to explore—to simplify our analysis and minimize risk we ignore any information about those humans and treat the responses for each item as, effectively, an anonymous sample.

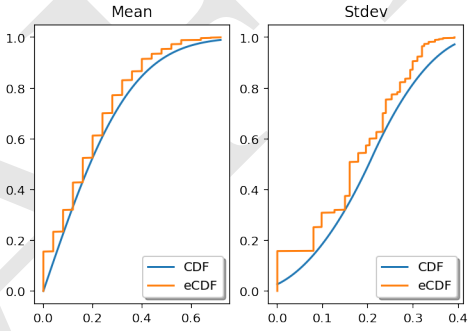
The first data was taken from SemEval-2024 Task 11: Learning with Disagreements (LeWiDi) (Leonardelli et al., 2023). We chose this dataset because is is among the very few we could find that retained individual, i.e., "disaggregated," rater responses; most datasets contain some aggregation of the responses, such as the mean (but not the variance), or the plurality, etc. The **MultiDomain Agreement** (Leonardelli et al., 2021) dataset contains tweets about Black Lives Matter, the US 2020 presidential election, and COVID-19 annotated for offensiveness. The test set has 3057 items annotated by 5 raters each. We fit the means and standard deviations of the item responses to *truncated* normal distributions with $(\mu = -0.5, \sigma = 1)$ and $(\mu = -0.3923, \sigma = 0.8502)$, respectively. This dataset does not appear to be publically available other than from the LeWiDi github site. The author of this dataset is also a co-author of (Leonardelli et al., 2023). Instructions for directly obtaining the dataset from the author are available at `https://github.com/dhfbk/annotators-agreement-dataset`.

The **Stanford Toxicity** dataset (Kumar et al., 2021) was also used in (Wein et al., 2023). It contains 107,620 items annotated by 5 raters each with ratings on a 5-point Likert scale: not/slightly/moderately/very/extremely toxic. We use the same distributions as they do, namely, a folded normal with $(\mu = 0.19, \sigma = 0.11)$ for the means and a triangular distribution with $(a = -0.05, b = 0.21, c = 0.45)$ for the standard deviations. The data is available at `https://data.esrg.stanford.edu/study/toxicity-perspectives`. It is encrypted, but the website gives instructions for how to decrypt it. There is no published license.

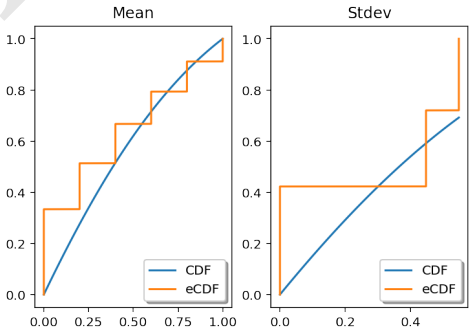
## 6 Results

We mainly used the following metrics in experiments:

- *Mean absolute error difference* (MAE). The distances (errors) from the per-item mean gold response to machine response averaged over the items: $\Gamma_{\text{MAE}}(A, B, G) = \frac{1}{N}\sum_i^N \left( \left| \frac{1}{K}\sum_j^K B_{ij} - \frac{1}{K}\sum_j^K G_{ij} \right| - \left| \frac{1}{K}\sum_j^K A_{ij} - \frac{1}{K}\sum_j^K G_{ij} \right| \right)$
- *Item-wise wins* (Wins). The fraction of items in the test set for which the absolute error of A is smaller than B: $\Gamma_{\text{Wins}}(A, B, G) = \sum_{i=1}^N \mathbf{1}_<(|\overline{A_i} - \overline{G_i}|, |\overline{B_i} - \overline{G_i}|)/N$
- *Mean EMD difference* (MEMD). The Earth mover's distance for each item between the system and the gold standard responses, and then take the mean of those item-wise EMDs: $\Gamma_{\text{MEMD}}(A, B, G) = \sum_{i=1}^N \left(\text{EMD}(B_i, G_i) - \text{EMD}(A_i, G_i)\right)/N$



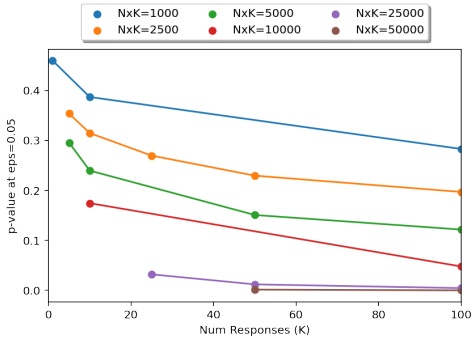

(a) Stanford Toxicity



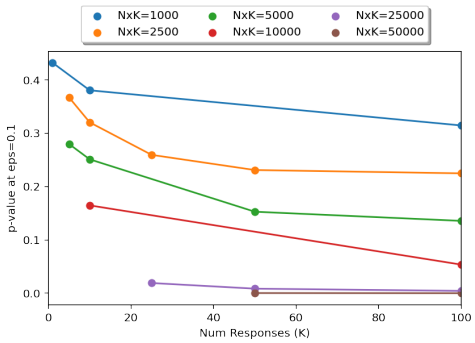

(b) MultiDomain Agreement

Figure 1: Empirical CDFs of item-level response means and standard deviations in (a) the Stanford Toxicity dataset vs clipped, folded normal CDF with $\langle \mu = 0.19, \sigma = 0.11 \rangle$ and clipped triangular distribution CDF with $\langle a = -0.05, b = 0.21, c = 0.45 \rangle$, respectively; and (b) the MultiDomain-Agreement dataset vs truncated normal CDF with $\langle \mu = -0.5, \sigma = 1 \rangle$ and truncated normal CDF with $\langle \mu = -0.3923, \sigma = 0.8502 \rangle$, respectively.

Upon publication we will release to the public the code used to run our experiments. We used

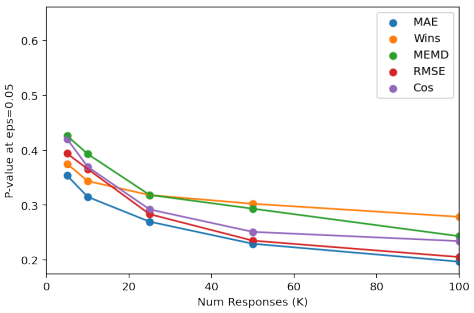



(a) Toxicity ($\epsilon = 0.05$)

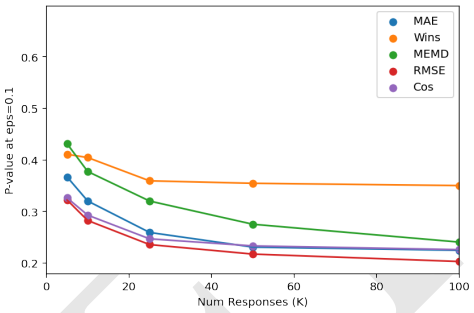



(b) MultiDomain ($\epsilon = 0.1$)

Figure 2: $p$-value vs $K$ with $\Gamma_{\text{MAE}}$ at various $N \times K$




(a) Toxicity ($\epsilon = 0.05$)




(b) MultiDomain ($\epsilon = 0.1$)

Figure 4: $p$-value vs $K$ with a fixed budget $N \times K = 2500$ for various metrics

the python libraries numpy, pandas, and scipy, versions 2.2.3, 2.2.1, and 1.13.1, respectively. Our experiments took various times to running, with the longest experiments (producing any of the points in our figures) running approximately nine hours.

Figure 2 demonstrates that trading off items for responses is beneficial at a wide range of $(N \times K)$ values, with $p$-value decreasing as $K$ increases. (The benefit of increasing $K$ is strikingly more apparent when viewing $p$-values vs $K$ with a fixed $N$ but we omit these graphs for brevity.) Here $\Gamma_{\text{MAE}}$ was used with distortion $\epsilon = 0.1$ but similar trends were observed using other metrics, amounts of distortion, as well as different datasets. There is indeed a point where trading $N$ for $K$ is beneficial for statistical significance: in this case, the curves hit an inflection point before $K = 500$; see Figure 3.

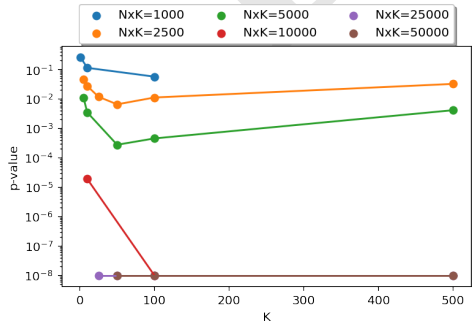


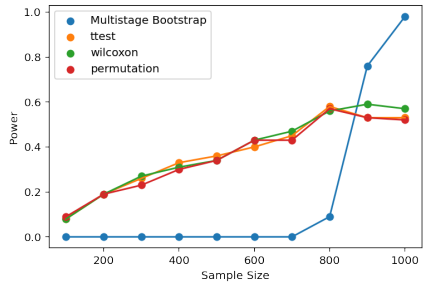

Figure 3: $p$-value vs $K$ with $\Gamma_{\text{MAE}}$ at various $N \times K$ for Toxicity at log-scale on the y-axis

Figure 4 graphs $p$-value as a function of number of responses at $\epsilon = 0.1$, where number of items varies such that $N \times K = 2500$, and demonstrates a similar trend across five different metrics.

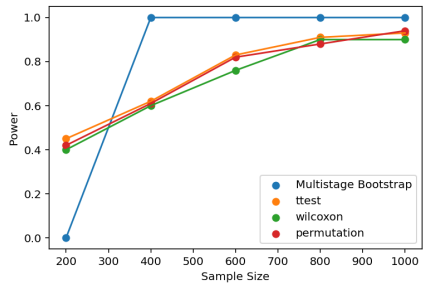
### 6.1 Power Analysis

Figure 5 demonstrates greater statistical power for Multistage Bootstrap as number of items ("sample size") increases, achieving a power of 90% (i.e., probability of not rejecting the null hypothesis when it's false) before the other tests. We used $\alpha = 0.05$ as the significance level for power calculation (i.e., the data is inconsistent with the null hypothesis at least 95% of the time).

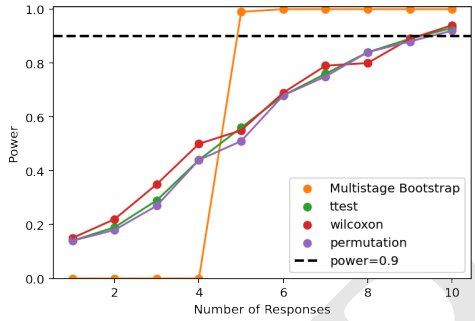
For the baseline (paired) statistical tests, the mean response of each item was pre-computed for Model A, Model B and for "gold" G, resulting in $\bar{a}_i, \bar{b}_i, \bar{g}_i$, respectively, for each item $i$. The baseline tests then consider the null hypothesis that the distributions across the items of $|\bar{a}_i - \bar{g}_i|$ and $|\bar{b}_i - \bar{g}_i|$ are the same in the case of the permutation test, or have the same center in the case of Welch's t-




(a) varying $N$ with $K = 5$




(b) varying $N$ with $K = 10$




(c) varying $K$ with $N = 1000$

Figure 5: Power Analysis of Toxicity data ($\epsilon = 0.1$)

test and the Wilcoxon signed-rank test. In contrast, our Multistage Bootstrap resamples both the set of items and, for each item, the set of responses once per iteration, hence considering the disaggregated distribution of responses.

## 7 Discussion

Our results confirm our predictions indicate that the number of raters and items do have a notable impact on $p$-value estimation, to different degrees depending on the metric. $\Gamma_{\text{Wins}}$ provides a discrete decision for each *item*, counting those decisions (i.e."wins") across the test set and normalizing by the number of items. $\Gamma_{\text{Wins}}$ is also presented as a meta-metric of sorts: it can use any item-level metric, with absolute error being used here, and requires both model's predictions as input in order to directly compare their predictions at the item level.

In general, increasing $N$ (number of test set items) increases the statistical power of any measurement by simply providing more scores to base the final metric score on. The more scores there are, the more stable the variance across simulation runs will be, and the lower the $p$-value. All examined metrics respond well to increasing $N$.

Increasing $K$ (number of responses per item) increases the statistical power of each *item level aggregate*. As $K$ increases, the lower the variance of an individual item's aggregate will be across simulation runs, thereby lowering the $p$-value. All tested metrics also respond well to increasing $K$.

The difference between the metrics lies in the way the item-level scores are used. For Wins, which responds better to increasing $N$, the $A$'s and $B$'s item-level scores are directly compared. In each run, these item-level scores will vary, but in many cases that variance won't change the pairwise comparison. For example, if $A_i$'s metric score is 0.10 and $B_i$'s is 0.12 on the first simulation, a win is recorded for $A$. In the next simulation, if the scores are 0.11 and 0.12, respectively, this score change does not change the Win, as $A_i$'s score is still lower. This indicates the item-level variance in the discrete win decision is far lower than the score variance - so adding more responses is less likely to further reduce the variance than adding items.

By contrast, for $\Gamma_{\text{MAE}}$ and $\Gamma_{\text{MEMD}}$, any changes in item-level metric scores do impact the variance, both at the item and test-set level. Since the item-level scores come from the response distribution, adding more responses stabilizes the simulated distributions under repeated test set generation, reducing the metric variance across simulations and lowering the $p$-value.

The implications of these results are that the item/response trade-off should be handled differently depending on the metric itself, and the demands on number of raters and items are high for all metrics in order to provide statistical guarantees.

## 8 Conclusion

In this work, we experimented with simulated data in order to examine the trade-off between number of items and number of annotations per item (aka responses) necessary to compare two systems against human judgments with statistical significance ($p < 0.05$). As expected, we see that when two systems are more similar in performance,

a greater number of annotations is required to achieve significance on their comparison. Further, the metric itself affects the utility of an increase in either items or responses.

These results suggest that current evaluation practices are not sufficient to confidently assess two systems' performance against gold judgments, as using 25,000-50,000 annotations in a test set is rarely seen. Even when using 1000 items, at least 25 raters are needed for systems to achieve significance with MAE.

Additionally, we found that the trade-off between number of items and number of responses per item, depended on the metric. For two of our tested metrics, MAE and mean EMD, adding more responses than items is a more optimal division to achieve lower $p$-values. For the Wins metric, the opposite is true: more items and fewer responses per item lead to lower $p$-values. Still, in all cases for all metrics, increasing the total number of responses consistently lowers $p$-values, and thereby increases the sensitivity of the evaluation instrument.

For real-world data, we actually found MAE to be more sensitive than Wins, and we speculate that this may be due to the discrete response domains in the real-world data, compared to the real-valued responses in the synthetic data.

## 9 Limitations

The effectiveness of (Wein et al., 2023)'s simulator depends on how well the probabilistic models capture realistic distributions of responses over items. Although we used rigorous methods to fit the parameters of these distributions to our datasets, our choice of distribution family to use for each dataset was based on visual inspection of the data. Given more datasets with disaggregated responses we hope in future work to develop rigorous methods for model selection. However, the dearth of such publicly-available datasets impedes progress in this direction. One key limitation future work will address is that we treat the responses as independent from item-to-item, when in reality responses usually depend on which human annotator or instance of a model produced the response. Hypothesis testing such as that described here is not a comprehensive measure of data quality; it only estimates the likelihood of sampling error. It does not account for sampling bias leading to data that is not representative of the sampling distribution.

The simulator is only intended to capture the complexity of the annotations. It is not intended to capture the complexity of real model predictions but rather to compare a near-perfect model, $A$, against a version, $B$, that has been perturbed by a controlled amount via a variance parameter. In practice, this functions as an approximate bound the model response variance.

Otherwise, we have taken precautions to avoid common "$p$-hacking" pitfalls, such as that the null hypothesis and significance threshold $\alpha$ are independent of the dataset. We attempt to avoid *optional stopping* by performing power analysis.

While the distribution of responses depend on each item, we do not assume a fixed correspondence between ratings and raters. This assumption is valid, for example, with a large rating pool where each rater annotates at most one item. Therefore, there is no meaningful ordering of the responses within each item. For convenience, we use the term "matrix" for what is really a sequence of multisets. Modeling the dependence of annotations from the same annotators across multiple items is something we chose to ignore in this paper so as not to distract from its main focus on the impact of response variance on hypothesis testing.

### Ethical considerations

The paper focuses on a method to ensure that enough data is collected during testing to ensure that large enough observed differences between the performance of two machines on the data are significant. While such analysis can ensure that experiment results are meaningful and replicable, p-values have a tendency to be used more than they are understood. It is important to understand what p-values guarantee and what the limitations of our, or any other particular NHST framework, are. Misinterpreting the analysis can lead dishonest or misleading claims about the reliability of the data for testing.

## References


R.B. Bausell and Y.F. Li. 2002. *Power Analysis for Experimental Research: A Practical Guide for the Biological, Medical and Social Sciences*. Cambridge University Press.

Hung Chau, Saeid Balaneshin, Kai Liu, and Ondrej Linda. 2020. Understanding the tradeoff between cost and quality of expert annotations for keyphrase extraction. In *Proceedings of the 14th Linguistic Annotation Workshop*, pages 74–86, Barcelona, Spain. Association for Computational Linguistics.

Paul R Daugherty and H James Wilson. 2018. *Human+ machine: Reimagining work in the age of AI*. Harvard Business Press.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2021. A statistical analysis of summarization evaluation metrics using resampling methods. *Transactions of the Association for Computational Linguistics*, 9:1132–1146.

Thomas G Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923.

Rotem Dror, Lotem Peled-Cohen, Segev Shlomov, and Roi Reichart. 2020. Statistical significance testing for natural language processing. *Synthesis Lectures on Human Language Technologies*, 13(2):1–116.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of nlp leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA. PMLR.

David Goldberg, Andrew Trotman, Xiao Wang, Wei Min, and Zongru Wan. 2018. Further insights on drawing sound conclusions from noisy judgments. *ACM Trans. Inf. Syst.*, 36(4).

Odd Erik Gundersen. 2020. The Reproducibility Crisis Is Real. *AI Magazine*, 41(3):103–106.

Michael Hagmann, Philipp Meier, and Stefan Riezler. 2023. Towards inferential reproducibility of machine learning research. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Jose Hernandez-Orallo. 2020. Ai evaluation: On broken yardsticks and measurement scales. In *Workshop on Evaluating Evaluation of Ai Systems at AAAI*.

Christopher M Homan, Shira Wein, Chris Welty, and Lora Aroyo. 2024. How many raters do you need? power analysis for foundation models. In *I Can't Believe It's Not Better Workshop: Failure Modes in the Age of Foundation Models*.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. 2021. Designing toxic content classification for a diversity of perspectives. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 299–318. USENIX Association.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Elisa Leonardelli, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, Alexandra Uma, and Massimo Poesio. 2023. SemEval-2023 task 11: Learning with disagreements (LeWiDi). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2304–2318, Toronto, Canada. Association for Computational Linguistics.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. *arXiv preprint arXiv:2109.13563*.

Christopher H. Lin, Mausam, and Daniel S. Weld. 2014. To re(label), or not to re(label). In *HCOMP 2014*.

Rachel Longjohn, Giri Gopalan, and Emily Casleton. 2025. Statistical uncertainty quantification for aggregate performance metrics in machine learning benchmarks.

Zeinab Mashreghi, David Haziza, and Christian Léger. 2016. A survey of bootstrap methods in finite population sampling. *Statistics Surveys*, 10(none):1 – 52.

Art B. Owen. 2007. The pigeonhole bootstrap. *The Annals of Applied Statistics*, 1(2):386 – 411.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.

M Poesio R Artstein. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.

Inioluwa Deborah Raji, Emily M Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. Ai and the everything in the whole wide world benchmark.

Stefan Riezler and Michael Hagmann. 2021. *Validity, Reliability, and Significance: Empirical Methods for NLP and Data Science*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P Lalor, Robin Jia, and Jordan Boyd-Graber. 2021. Evaluation examples are not equally informative: How should that change nlp leaderboards? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

Sanjif Shanmugavelu, Mathieu Taillefumier, Christopher Culver, Oscar R. Hernandez, Mark Coletti, and Ada Sedova. 2024. Impacts of floating-point non-associativity on reproducibility for HPC and deep learning applications. *CoRR*, abs/2408.05148.

Noam Shazeer, *Azalia Mirhoseini, *Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*.

Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. 2008. 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, page 614–622, New York, NY, USA. Association for Computing Machinery.

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

Anders Søgaard. 2013. Estimating effect size across datasets. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 607–611, Atlanta, Georgia. Association for Computational Linguistics.

Anders Søgaard, Anders Johannsen, Barbara Plank, Dirk Hovy, and Hector Martínez Alonso. 2014. What's in a p-value in NLP? In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 1–10, Ann Arbor, Michigan. Association for Computational Linguistics.

Student. 1908. The probable error of a mean. *Biometrika*, pages 1–25.

Piotr Szymański and Kyle Gorman. 2020. Is the best better? Bayesian statistical model comparison for natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2203–2212, Online. Association for Computational Linguistics.

Aobo Wang, Cong Duy Vu Hoang, and Min-Yen Kan. 2013. Perspectives on crowdsourcing annotations for natural language processing. *Language resources and evaluation*, 47:9–31.

Shira Wein, Christopher Homan, Lora Aroyo, and Chris Welty. 2023. Follow the leader(board) with confidence: Estimating p-values from a single test set with item and response variance. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3138–3161, Toronto, Canada. Association for Computational Linguistics.

Peter Welinder and Pietro Perona. 2010. 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. IEEE.