With Little Power Comes Great Responsibility

Dallas Card[1] **Peter Henderson**[1] **Urvashi Khandelwal**[1] **Robin Jia**[1]
Kyle Mahowald[2] **Dan Jurafsky**[1]
[1]Stanford University, Stanford, CA
[2]University of California Santa Barbara, Santa Barbara, CA
dcard@stanford.edu, phend@stanford.edu,
urvashik@stanford.edu, robinjia@stanford.edu,
mahowald@ucsb.edu, jurafsky@stanford.edu

Abstract

Despite its importance to experimental design, statistical power (the probability that, given a real effect, an experiment will reject the null hypothesis) has largely been ignored by the NLP community. Underpowered experiments make it more difficult to discern the difference between statistical noise and meaningful model improvements, and increase the chances of exaggerated findings. By meta-analyzing a set of existing NLP papers and datasets, we characterize typical power for a variety of settings and conclude that underpowered experiments are common in the NLP literature. In particular, for several tasks in the popular GLUE benchmark, small test sets mean that most attempted comparisons to state of the art models will not be adequately powered. Similarly, based on reasonable assumptions, we find that the most typical experimental design for human rating studies will be underpowered to detect small model differences, of the sort that are frequently studied. For machine translation, we find that typical test sets of 2000 sentences have approximately 75% power to detect differences of 1 BLEU point. To improve the situation going forward, we give an overview of best practices for power analysis in NLP and release a series of notebooks to assist with future power analyses.[1]

1 Introduction

Despite its importance to empirical evaluation, relatively little attention has been paid to statistical power in NLP. In particular, *if it is the case that typical experiments in NLP are underpowered*, not only would we expect many meaningful improvements to go undetected, we would also expect many apparently significant differences to be exaggerated (Gelman and Carlin, 2014). In this paper, we build on past work calling for greater rigor in evaluation (McCoy et al., 2019; Azer et al., 2020), including the need for careful hypothesis testing (Koehn, 2004; Berg-Kirkpatrick et al., 2012; Søgaard et al., 2014; Dror et al., 2018), and show why and how power matters to NLP, addressing challenges unique to this domain.

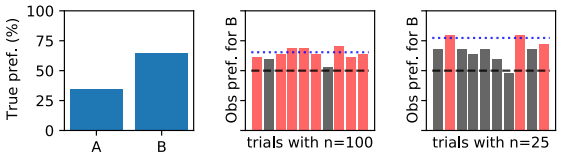

Figure 1: Cartoon example of statistical power in comparing two models: 65% of all people in the population always prefer system B (left). A comparison using a sample of 100 people would be well-powered (middle): over 80% of such samples will show a significant difference (plotted in red) from the null hypothesis that the models are equally good (dashed line). In samples of 25 people (right), far fewer tests will be significant (power $\approx 30\%$). Note that the observed mean of significant findings (dotted line) slightly overestimates the true proportion that prefer system B when $n = 100$ and more severely overestimates it when $n = 25$.

Roughly speaking, power is the probability that a statistical test will successfully detect a *true effect*. As an illustrative example, imagine comparing two dialog systems (see Figure 1). We want to know if people tend to prefer one system over the other. To test this, we will need multiple people to evaluate the systems. But how many? Once we have collected data, a statistical *test* will tell us if we can reject the null hypothesis the systems are equally good. Assuming the systems are not identical, statistical *power* is the probability that the experiment will return a significant result (or equivalently, it is one minus the probability of failing to detect the difference as significant). Although we don't know the magnitude of this difference, *power analysis* helps to estimate how much power an experiment

[1]https://github.com/dallascard/NLP-power-analysis

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 9263–9274,
November 16–20, 2020.

will have under various assumptions.

Power depends on multiple factors, including the statistical test used, the significance threshold, true effect size, variance, and sample size. All else being equal, experiments with larger samples will have greater power than smaller samples, as shown in Figure 1. Similarly, larger effects and those with less variance are easier to detect, and therefore require fewer samples for equivalent power. Importantly, note that if we *do* find a significant difference, this does *not* imply that the experiment had high power.[2]

Proceeding with a test that is *underpowered* (i.e., too few subjects or items; often taken to mean less than 80% power; Cohen, 1962) means that one is less likely to be able to draw any useful statistical conclusion from the experiment, and has contributed, in part, to the replication crisis in other fields (Button et al., 2013; Szucs and Ioannidis, 2017; Ioannidis et al., 2017). Routinely running experiments with low statistical power undermines the scientific enterprise. Not only will true effects go undetected; when significant effects are found, they are likely to be noisier and have lower positive predictive value (Button et al., 2013).

Moreover, significant findings from underpowered experiments are more likely to exaggerate or reverse the true effect – so-called Type-M (magnitude) and Type-S (sign) errors, respectively (Gelman and Carlin, 2014). This problem can lead to systematic distortions in the literature if only significant findings are published, especially if these results are based on underpowered experiments (Scargle, 1999). The effect of Type-M error can be seen in Figure 1; significant differences are less likely to be found in smaller samples (right), but among those tests that are significant, the observed difference will tend to exaggerate the true difference (left) by more than a larger sample (middle). For further discussion of Type-M and Type-S errors, please refer to Appendix B.

Here, we investigate how these issues affect NLP. Although retrospective analysis of power involves challenges, we present evidence that underpowered experiments are widespread in NLP research. Among human evaluations, we find most experimental designs involve too few items and/or raters to detect small effects (§5). For comparing models in terms of accuracy, we find that some widely used benchmark datasets, including MRPC and SST-2, are now too small to be able to properly measure future progress against top performing models (§3). We also introduce a novel approach to power analysis for machine translation and characterize power in experiments testing for differences in BLEU (§4). Finally, a survey of recent papers reveals a general lack of statistical evaluation and a dearth of detailed reporting (§5.1).

To improve future practice, we suggest broader adoption of power analyses prior to evaluation, provide guidance on running power analyses in NLP, and release a series of notebooks for this purpose.

2 Power Analysis for NLP

Because most NLP tasks do not take the form of standard experiments in other sciences (Kraemer and Blasey, 2015; Westfall et al., 2014), it is non-trivial to run power analyses for many tasks of interest. While we cannot cover every scenario, we present here a generalizable, simulation-based approach to power analysis, along with three sample applications, which can be extended as necessary. Such an approach is modular, reusable, and transparent, and encourages planning of analyses in advance of data collection.

Every power analysis requires assumptions, and there is not likely to be a single correct approach. Rather, the point is to make one's assumptions explicit, and include enough detail so as to account for whatever is likely to be observed. By using reasonable assumptions, one can help to ensure that one's experiment is sufficiently well-powered, In the case of NLP, this means that one recruits enough subjects, collects enough ratings, or uses a large enough test set.

The general procedure we suggest for power analysis is described in detail in Figure 2. At a high level, the idea is to estimate power by running simulations. Recall that power is the probability of detecting a true effect, conditional on the experimental setting (effect size, variance, etc.) and significance threshold. Thus, if one can translate these assumptions into a process for generating simulated data, we can estimate power by generating many simulated datasets using assumed or estimated parameter values, running each sample through a significance test, and reporting the proportion that are found to be significant.

[2]Using the observed outcome from a single experiment to compute power falls into the trap of post-hoc power analysis and is not recommended. For additional background on statistical power, power analysis, null-hypothesis significance testing, and post-hoc analysis, please refer to Appendix A.

Define a generative process $G(n, e^*, \mathbf{h})$ parameterized by number of items, n, hypothesized effect e^* for the statistic of interest E, and other relevant parameters $\mathbf{h}$ (e.g., variance). Also choose a statistical test $T(\mathcal{D})$, which returns a p-value p when performed on data $\mathcal{D}$ sampled from $G(n, e^*, \mathbf{h})$. Finally, choose the size of the dataset to be sampled, n, significance threshold, α, and number of repetitions, r.

1. For i in range(r):
   - sample a dataset of size n, $\mathcal{D}_i \sim G(n, e^*, \mathbf{h})$
   - compute the effect of interest on this sample, $e_i = E(\mathcal{D}_i)$
   - also compute a p-value according to the test of interest: $p_i = T(D_i)$
2. Power $\approx \frac{1}{r} \sum(\mathbb{I}[p_i \leq \alpha] \cdot \mathbb{I}[\text{sign}(e_i) = \text{sign}(e^*)])$

Figure 2: An algorithm for power analysis by simulation. For the example of comparing two systems presented in Figure 1, e^* is the assumed overall proportion of people who prefer system B, relative to the null hypothesis, $p = 0.5$, $G(n, e^*, \mathbf{h})$ is simply Binomial$(n, 0.5 + e^*)$, while e_i is the observed proportion of people who prefer system B in sample i, again relative to 0.5. For extensions to estimate Type-M and Type-S error, see Appendix B.

The key to generalizing this approach is to begin with the end in mind. In particular, if one plans to test for a difference between models, one needs to choose the statistical test that will be used. That test will determine the level of detail required in the generative process for simulating data.

To return to the opening example of evaluating dialog systems, we want to test if people prefer one system over the other (Ai et al., 2007). If we ignore the nuances of human preference for now (but see §5 for a more nuanced approach), and simply assume that each person either prefers system A or system B, the only assumption we need to make for a power analysis in this setting is the proportion of people in the population who prefer system B. We can then simulate samples of n people (each of whom independently has the same probability of preferring system B) as a draw from a binomial distribution, and repeat this thousands of times.[3] For each sample, we then test whether the proportion of people who prefer system B is significantly different from 0.5. The estimated power of this experiment would thus be the proportion of simulated differences that are found to be significant.[4]

The most difficult part of power analyses is estimating the relevant quantities, such as the *true* proportion of people that prefer system B. Note, however, that one can always compute what power would be for a range of possible values, and indeed, this is the recommended procedure. For estimating the relevant parameters within an NLP context, we will primarily rely on data from the literature, measurements on validation data, and estimates from external datasets (see §3.2). However, where appropriate, pilot studies may also be informative.

In the remainder of this paper, we consider three scenarios of interest in depth, and assess the state of power in the NLP literature for each.

3 Comparing Models on Accuracy

It is common in NLP research to look for models which improve over state of the art (SOTA) on various benchmarks. However, an important but rarely asked question is, *can these benchmarks support the kinds of comparisons we want to make?* Many have emphasized the need for proper significance testing to avoid spurious findings, but if an experiment's test set is small, the minimum detectable effect (MDE) size may be large: only large improvements will yield sufficiently powered comparisons (i.e., $\geq 80\%$ power). If an experiment is badly underpowered, it cannot provide useful evidence that one model achieves slightly better performance than another for the underlying data distribution. Reliance on such evidence risks leading to over-confidence about the relative ranking of various models. As we show in §3.3, there is legitimate reason to be concerned about this in the case of certain widely used benchmarks.

3.1 Significance test for comparing classifiers

The standard statistical test for comparing classifiers on paired data is McNemar's test (Dietterich, 1998; Dror et al., 2018), which uses the numbers of items where the models disagree (i.e., the off-diagonal elements in Table 1).[5] McNemar's test assesses whether $\chi^2 = \frac{(p_{10} - p_{01})^2}{p_{10} + p_{01}}$ is significant, and if so, rejects the null hypothesis that the distributions are the same.

[3] We don't need to address variance in this scenario, as the variance of a binomial distribution is a function of its mean.

[4] More direct solutions are available for some settings, including this one (see Appendix E.5), but we describe it using the generic approach from Figure 2 for the purpose of illustration. For all cases examined in this paper, simulations take only minutes on a laptop.

[5] Unpaired data (i.e., if two models are evaluated on different data drawn from the same distribution) requires a different approach, such as using a binomial test. See Appendix E.5 for extended discussion.

	M1 correct	M1 incorrect
M2 correct	both correct	only M2 correct
M2 incorrect	only M1 correct	both incorrect

Table 1: A contingency table representing the distribution of possible outcomes for two models (M1 and M2).

Thus, for McNemar's test, the relevant data generating process for simulations can be specified in terms of the expected difference in accuracy between the models, Δ_{acc}, and P_a, the expected proportion of examples for which the models will have the same outcome (i.e., both correct or both incorrect). From these we can compute the expected proportions of examples on which only one model is correct (i.e., the off-diagonals in Table 1), and estimate power via the algorithm in Figure 2. Figure 3 illustrates how power increases with increased sample size, effect size, and agreement rate.[6]

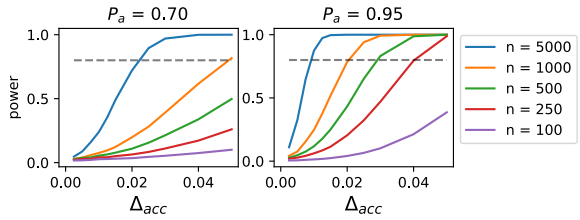

Figure 3: Power for comparing two classifiers on accuracy using paired data depends on the size of the test set (n), the expected agreement (P_a), and the expected difference in accuracy (Δ_{acc}). The dashed line shows 80% power, often taken to be a minimal requirement.

3.2 Estimating parameters

In order to estimate the required parameters (P_a and Δ_{acc}), we consider three options: (1) use results on validation (dev) data; (2) fit a regression based on historical data; (3) use middle-of-the-road assumptions when lacking other information. Using these methods, we can then estimate power or calculate the smallest effect that can be detected with 80% power at $\alpha = 0.05$ (or other thresholds). Both to illustrate this process, and to provide guidance for future work, we demonstrate these approaches below using data from two widely-used datasets for evaluating NLP models: SQuAD 2.0 (Rajpurkar et al., 2016, 2018) and the GLUE benchmark (Wang et al., 2018).

[6]Corresponding plots showing Type-M and Type-S error (Gelman and Carlin, 2014) are in Appendix B. To walk through a numerical example, see Appendix C. For an interactive example, see the accompanying online notebooks.

Using validation results: To the extent that we expect performance on test data to match performance on validation data (i.e., in the absence of domain shift), *paired* performance on validation data (i.e., difference in accuracy and agreement rate) provides one method for estimating power when comparing against a baseline model.

To illustrate this, from the authors of SQuAD 2.0, we obtain the pairwise agreement rates between all models submitted to the leaderboard on both validation and test data. We find a very strong correlation between validation and test for both pairwise accuracy differences (Δ_{acc}) and agreement rates (P_a) ($r = 0.99$ for both, as shown in Figure 9 in Appendix D, with results on validation data included in the accompanying online materials), suggesting we can use paired predictions on validation data for power calculations when we have access to the predictions from both models. Note that this approach assumes that the dev and test data have been drawn from the same distribution, and that dev performance has not been artificially inflated (such as by training on validation data directly).

Using historical data: When one does not have access to the baseline model or an informative prior, one can make use of historical trends. That is, we can try to estimate what a typical improvement will look like, given the current state of the art (SOTA). To illustrate this approach, we collect reported results for both SQuAD 2.0 and GLUE, and fit regressions to estimate Δ_{acc} and P_a. Given these parameters, we can assess the likely power and MDE for a typical model improvement against a given baseline accuracy level.

To fit a regression to predict typical improvements to SOTA, we gather data from GLUE papers and manually label 119 accuracy comparisons and 57 claims of improvement (as denoted by bolding of a result and a claim of SOTA in text) across 14 papers (selected as being at or above the BERT score on the GLUE leaderboard with an accompanying paper). In regressing Δ_{acc} on baseline accuracy and task, we achieve an $R^2 = 0.69$, which is not a perfect fit, but still provides a prior on likely effect size. Similarly, we achieve an $R^2 = 0.67$ when fitting a regression to SOTA improvements on the SQuAD 2.0 leaderboard (selected as being a significant improvement in time-ordered submissions). See Appendix E.2.1 for more details.

To assess power for McNemar's test, we must also fit a regression predicting the expected overlap

between the models (P_a). To fit such a regression, from GLUE authors we obtain the model test set predictions on all tasks from a set of 10 high-performing models, which allows us to measure the extent to which their predictions overlap with each other. Using GLUE tasks which measure accuracy, we regress P_a on baseline accuracy and Δ_{acc}, and achieve an R^2 of 0.97.[7] Repeating this for SQuAD 2.0, we get an R^2 of 0.94. See Appendix E.2 for regression coefficients and additional details.

Typical improvements on popular tasks tend to be small (see mean improvements in Table 2). Except for rare transformative work, such as BERT (Devlin et al., 2019), it is generally difficult to do *much* better than a previous SOTA and thus improvements are likely to follow a trend, which is why we are able to use historical data as a guide. In cases where such data is not available or cannot be trusted, other methods are necessary.

No prior: If no informative prior is available and the baseline model or can't be used for comparison on a validation set, then we must fall back on middle of the road assumptions. Lachenbruch (1992) provides a suggested default prior, and we find that MDEs using this method are very similar to those found by using the regression based approach. Appendix E.3 provides more details, and Table 9 in the appendix presents the comparison.

3.3 Assessing power in the literature

Using the regression-based approach of estimating Δ_{acc} and P_a described above, we estimate the MDE for each individual accuracy-based GLUE task in comparison to current SOTA, and report the average effect size of results which claimed improvements. Table 2 summarizes these results, showing for each dataset the size of the test set, the accuracy of the best performing model on each task at the time of writing, the estimated MDE to have 80% power using our regression to predict overlap (P_a), and the average reported difference from their respective baselines.

As can be seen in Table 2, the mean reported effect size ($|\Delta_{acc}|$) is well below the estimated MDE for the three smallest test sets – WNLI, MRPC, and SST-2. Because this mean is based on models comparing to even weaker baselines, we would expect most future improvements to be even smaller. Thus, most future experiments involving these three datasets *will not have adequate power* to test for improvements over the current SOTA in the way that they are routinely used. Moreover, alternative analyses give *even more pessimistic* estimates of likely improvements relative to MDE, as described in Appendix E.4. If an experiment does show significant improvement on a dataset such as MRPC, the potential for Type-M error should make us skeptical that this improvement will generalize to new data from the same domain.

Dataset	Size	SOTA (%)	Est. MDE (%)	$\lvert\Delta_{acc}\rvert$ (%)
WNLI	147	94.5	+5.26	+1.72
MRPC	1,725	92.0	+1.62	+0.63
SST-2	1,821	97.2	+1.02	+0.57
RTE	3,000	91.7	+1.23	+3.89
QNLI	5,463	97.5	+0.55	+1.31
MNLI-m	9,796	91.6	+0.67	+0.97
MNLI-mm	9,847	91.3	+0.68	+1.29
QQP	390,965	91.0	+0.11	+0.36
SQuAD 2.0	8,862	90.7	+0.56	+2.23†

Table 2: Estimated minimum detectable effect (MDE) using a regression-based estimate of likely agreement with leaderboard SOTA as of May 6th, 2020. $|\Delta_{acc}|$ is the average improvement over baseline per task among surveyed papers that claimed SOTA. For future comparisons, unless the expected improvement is larger than the estimated MDE, an experiment is unlikely to be adequately powered, and researchers should instead choose a different (larger) dataset. Note that this likely applies to the vast majority of experiments on WNLI, MRPC, and SST-2, based on recent trends. † indicates that the SQuAD 2.0 average was based on leaderboard improvements, which weren't necessarily reported in a publication. See Appendix E for full table and details.

While the above results are informative about future experiments, we would also ideally like to know about the power of past experiments. Most of the papers from which we collected results did not report a significance test on the test set. Here we estimate the expected power and predicted result of such a test using leave-one-out regressions, where we make a prediction for each reported improvement using all other reported model comparisons. This procedure reveals that **only 46% would have predicted adequate power** (using estimates for expected improvement and agreement), and **approximately 51% would have been significant** (based on estimated agreement and *reported* improvement). Approximately 80% of experiments with at least 80% power would also have been

[7] WNLI (Levesque et al., 2012), MRPC (Dolan and Brockett, 2005), SST-2 (Socher et al., 2013), RTE (Dagan et al., 2005; Bar-Haim et al., 2006; Giampiccolo et al., 2007; Bentivogli et al., 2009), QNLI (Rajpurkar et al., 2016) MNLI (Williams et al., 2018), and QQP (Iyer et al., 2017). For consideration of other metrics, see Appendix F.

found to be significant (37% of all comparisons).

In part because performance on many of these tasks is now so good, a large expected improvement is required in order for a new experiment to have 80% power, suggesting that larger test set sizes may be necessary to continue making well-powered claims of SOTA improvement on individual tasks. For any comparisons which are likely to be underpowered, we should refrain from placing much emphasis on obtaining small improvements over the previously reported best model. In extreme cases, such as MRPC and SST-2, it is worth considering whether it is time to retire these datasets as the basis for model comparison.[8]

4 Machine Translation

To show how our approach to power analysis can be applied to a more difficult setting, we consider automated evaluation of machine translation using BLEU scores (Papineni et al., 2002). As with accuracy, we would like to know what scale of improvements can be detected with reasonable power on typical test sets. This setting is more complicated because (1) BLEU is a corpus-level metric, rather than being averaged across instances, and (2) typical models are trained on vast amounts of parallel data, with little data available that has not been used in training, making it difficult to estimate variation in performance.

Significance testing for BLEU: To test for a significant difference between two MT models we use the randomization test, as recommended in Dror et al. (2018): given the paired output translations from both models, swap the outputs for a random subset of test examples and compute the resulting difference in BLEU. Repeating this thousands of times gives us a null distribution, which can be used to test the observed difference between models.

Generative process for simulations: If large amounts of untouched evaluation data were available, we could approach power analysis by simply evaluating BLEU score on many random subsets of n sentences, and computing the mean and variance of each system. Unfortunately, because MT depends on parallel text (most of which is used in training), evaluation data tends to be scarce. Instead, we introduce a generative process that can produce the necessary inputs for power analysis.

For intuition, note that if we swap the i^{th} pair of model outputs (as is done in the randomization test), leaving rest as they are, we change the difference in BLEU between models by a specific amount, δ_i, which we call the effect of making that swap. While these individual effects are not independent of each other due to the corpus-level nature of the metric, in practice, the sum of individual effects closely approximates the net effect of swapping entire subsets (see Figure 15 in Appendix G).

Based on analyzing several models and datasets, we find the typical distribution of these individual effects can be approximated using a mixture of a Delta distribution at zero, and a Laplace distribution (see Appendix G for details). Concretely, if we assume Δ_B is the expected difference in BLEU between two models on a dataset of n examples, and P_0 is the expected proportion of examples for which $\delta_i = 0$, we can simulate a dataset $\{\delta_i\}_{i=1}^n$ of n individual effects using the following process: with probability P_0, $\delta_i = 0$. With probability $1 - P_0$, $\delta_i \sim \text{Laplace}(\mu, b)$, where $\mu = \frac{-2 \cdot \Delta_B}{n(1-P_0)}$, $b = b_0/n$, and b_0 is a user-specified parameter that controls the variance, independent of the sample size. By construction, $\mathbb{E}[\sum_{i=1}^n \delta_i] = -2 \cdot \Delta_B$.[9]

Given this generative process, we can then estimate power using the Algorithm in Figure 2. On each iteration, draw a simulated dataset from the generative process, compute the observed difference between models as $\hat{\Delta}_B = -\frac{1}{2}\sum_{i=1}^n \delta_i$, and test if this is significantly different from zero using a modified randomization test, in which we assume that the net effect of swapping a subset of instances is simply the sum of the δ_i's in the subset. (Please see online materials for an interactive example).

Empirical estimates: In order to estimate reasonable values for the required parameters, we use several pretrained models from the FAIRSEQ library (Ott et al., 2019) for the WMT English-German translation task. We evaluate these models on the shared task test sets from 2016-2019 and compute BLEU scores using SACREBLEU (Post, 2018). Fitting a Delta-Laplace mixture to the effects of swapping individual output pairs, we estimate values for $\hat{P}_0$ and $\hat{b}_0$, reported in Table 3. (See also Figure 16 in Appendix G; code for computing estimates is provided in the online materials).

[8] It is also worth exploring power with respect to claims of improvement on multiple tasks with a single model (Demšar, 2006), rather than each task individually. We leave consideration of this as an interesting direction for future work.

[9] Note that swapping all n examples would reverse the model scores, equivalent to a net effect of $-2 \cdot \Delta_B$.

M1	M2	Test set	n	Δ_B	$\hat{P}_0$	$\hat{b}_0$
TF19*	TF18*	2019	2K	4.3	0.19	23.7
TF18	TF16	2018	3K	4.2	0.09	29.4
TF16	Conv17	2017	3K	1.3	0.12	22.5
TF16	Conv14	2016	3K	7.6	0.10	27.6

Table 3: Relevant parameters from four MT evaluations. TF are Transformer-based (Ott et al., 2018; Edunov et al., 2018; Ng et al., 2019) and Conv are Convolutional models (Gehring et al., 2017) from FAIRSEQ. Test sets are from WMT shared tasks for En-De translation. Δ_B is the reported difference in BLEU, whereas $\hat{P}_0$ and $\hat{b}_0$ are estimated. * indicates ensembles.

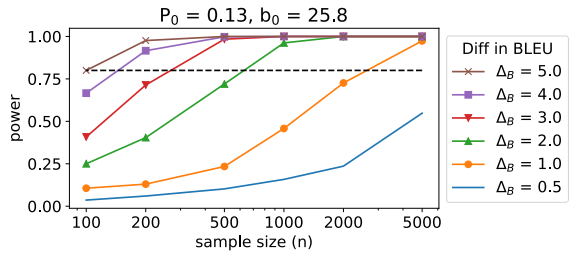

Figure 4: Power analysis for MT, showing how power increases with n and Δ_B, using an average of fitted values for P_0 and b_0. Based on this analysis, we expect that an experiment with a test set of 2000 sentences would have approximately 75% power to detect a difference of 1 BLEU point as significant. For additional plots, refer to Figure 17 in Appendix G.

While far from identical, the four comparisons, each representing different stages of model evolution, all produce similar estimates. Although these estimates are only based on a single language pair, the models and test sets are relatively diverse, and we expect that these estimates will generalize, though better estimates could be obtained by fitting this distribution to a new domain of interest.

Using these estimates, we can now characterize how much power test sets of different test set sizes (n) would have for a range of possible differences in BLEU (Δ_B). Figure 4 shows this for P_0 and b_0 set to the average of the observed values.[10] Based on this estimate, we conclude that for typical MT test sets of around 2,000 examples, an improvement of 1 BLEU point can likely be detected with approximately 75% power. As shown in Figure 4 this power level increases dramatically with sample size and effect size.

This analysis has served, in part, to show how a simulation-based approach to power analysis can be adapted to virtually any task. Additional work is required to test how well these specific parameter estimates will generalize, but the same process can easily be adapted to new language pairs. More generally, there would be great value in the MT community curating larger held-out test sets, both to validate this analysis, and for better powered future comparison.

[10]For a sensitivity analysis of how power varies under different assumptions for P_0 and b_0, please see Figure 17 in Appendix G.

5 Likert-Scale Human Evaluations

Tasks such as natural language generation are difficult to evaluate using automated methods; as such, human evaluations are central to NLP. Past work has reported great variation in how human evaluations are done (van der Lee et al., 2019). Therefore, we begin with a meta-analysis of a subset of human evaluation experiments from EMNLP 2019, which we then use as the basis for claims about the power of human evaluations in NLP more generally.

5.1 Meta-analysis

To characterize the state of human evaluation in NLP, we identified papers from the main session of EMNLP 2019 that made use of human evaluations (details in Appendix H.2). To generalize across studies, we restrict our analysis to Likert-scale comparisons, which was the most commonly reported type of evaluation. We extracted all cases where a new model was being compared to the best-performing baseline on one more metrics (117 comparisons from 41 papers) and normalized all ratings to be on a 0-1 scale.

One takeaway from this meta-analysis is that the reported effect sizes (that is, difference between the novel model and the best-performing baseline) vary widely (s.d. = .12 on a [0, 1] scale). Number of items tested is more consistent: 69% used 100 or fewer, and only 18% used over 200. But, as similarly found by van der Lee et al. (2019), many key details were not reported in this sample of experiments. Most commonly missing was number of ratings per item (34% of all experiments), followed by total number of workers (28%). For 7% of experiments, we could not determine the number of items tested. 57% of experiments collected 3 annotations per item, which was also the modal number of unique annotators. Thus, it is often difficult to ascertain, for any particular experiment, the details of the experimental setting that are necessary to evaluate the validity of the results.

Because the number of items rated was the most

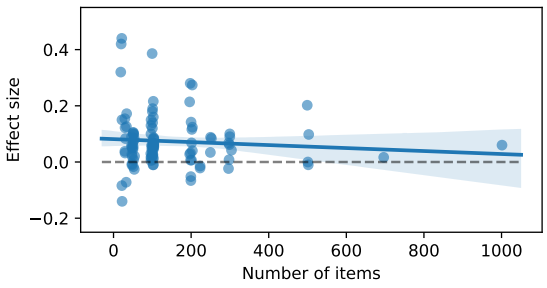

Figure 5: Scaled effect size vs. number of items from our EMNLP 2019 survey, showing higher variance in the smallest samples. There is a slight negative correlation, though it is not significant. As can be seen, most experiments are small ($n \leq 100$).

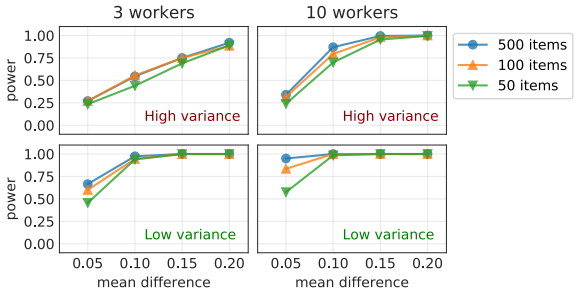

Figure 6: Using parameters estimated with mixed effects models from a high variance setting (top) and a low variance setting (bottom), the left panel shows simulated experiments with 3 workers annotating each item, the right panel shows an unusually high number of annotators per item (10 workers). Under typical assumptions, many common experimental settings (e.g., 3 workers and 100 items) are underpowered.

commonly reported, we use that as our proxy for sample size. Figure 5 shows scaled mean difference between models as a function of number of items. As expected, we see greater variance in effects with smaller samples since, with smaller samples, we expect greater noise. We also observe a slight negative correlation between effect size and sample size. That is, as sample size gets larger (and, thus, as estimates get more precise), the estimated effect size gets smaller. This trend is sometimes used as an indication of publication bias (censoring of null and opposite-direction effects) since, in a sample with no publication bias, the effect size should be independent of the sample size (Begg and Mazumdar, 1994). However, in our case, this correlation is not significant (Kendall's $\tau = -.07$, $p = .32$) and so it is difficult to draw strong conclusions.[11]

5.2 Power analysis for human Likert ratings

What kind of effect sizes can typical human evaluation experimental designs detect? As in previous sections, we can use simulations to explore how many annotators and/or instances should be used to have sufficient power.

Simulating human experiments is conceptually simple (e.g., m raters each rate n generated sentence on overall quality), but for realistic simulations, we need to consider variation in items (some generated sentences are better than others), and variation by rater (some raters use higher ratings and/or respond to different aspects of quality), as well as the overall difference in quality between models. A simulation which treated all workers as identical would fail to capture this variation, and hence might overestimate power (Barr et al., 2013).

[11] We exclude from this analysis two large negative effects with $N = 500$ which would exaggerate this correlation.

Unfortunately, details such as worker variance are rarely reported in published papers. To better characterize the typical variation in human evaluations, we rely on a convenience sample of several large datasets to estimate these parameters and use them in our simulations as a proxy for what we might observe in practice. Although focused on different tasks, all use a similar methodology, namely, getting many Likert-scale annotations per instance from many annotators and models (in some cases as many as 20 ratings per item).[12]

In order to extract estimates of these parameters for our simulations, we use hierarchical mixed-effects models, as used in psychology and other behavioral fields (Barr et al., 2013; Gelman and Hill, 2006). Such models incorporate variation in the quality of generated instances, annotator responses, and annotator sensitivity, and are recommended by van der Lee et al. (2019) for analyzing human evaluations. (We provide details in Appendix H.3 and include code for fitting such models as part of the online materials). Using this approach, we obtain an estimate of the relevant parameters from each of the large datasets. From these, we choose sets of parameters to be representative of experiments with high or low variance, with full results in Appendix H.3 (see Table 16 for parameter estimates).

As before, we then use these estimates to simulate data, assess significance on the simulated data (here using mixed effect regression), and compute power as a function of mean difference and sample

[12] We use publicly available or author-provided data from Hashimoto et al. (2019); Dathathri et al. (2020); Holtzman et al. (2020), and WMT19 (links in Appendix H.2).

size.[13] The resulting power estimates are shown in Figure 6, plotted in terms of effect size, sample size, and numbers of workers and items, for both the high and low variance scenarios. From this analysis, we highlight a few key takeaways:

- *Many human evaluation studies are likely underpowered:* Using the "high variance" parameters (which are typical of most of the datasets we used), the most common design at EMNLP 2019 (3 workers, 100 items) is underpowered unless the effect size is quite large (0.2 or higher on the [0, 1] scale).
- *Even with low variance, typical designs are underpowered to detect small effects:* Using our estimated parameters for the low variance setting, experiments will be underpowered to detect small effects (0.05 on the [0, 1] scale), unless an unusually large number of ratings per item are collected (10+ for 100 items).
- *Need for improved reporting:* Most human evaluations do not report enough detail to interpret the results. This could be drastically improved through basic power analyses, significance testing using mixed-effects models, and sharing of raw data.

Given our model estimates and simulations, we conclude that, in aggregate, many human evaluations are underpowered and would benefit from larger sample sizes, particularly by using more workers per item. Increased adoption of even approximate power calculations within the NLP community will promote thoughtful consideration of appropriate sample sizes and improve the reliability and replicability of results.

6 Overall Recommendations

- Power analyses should be done prior to evaluation when comparing against a baseline. If a comparison is likely to be underpowered, the pros and cons of running that evaluation should be carefully considered. Underpowered experiments do not provide convincing evidence of progress.
- For new datasets and shared tasks, the number of instances in the test will determine the minimum detectable effect size, and should be chosen accordingly.
- For tasks which no longer have adequate power to detect typical improvements (e.g., MRPC and SST-2), authors should consider expanding the test set or retiring the task.
- To facilitate future power calculation and significance tests, model owners should release final fine-tuned model checkpoints. Alternatively, leaderboard owners may wish to make validation set predictions from all submitted models publicly available.
- For human evaluations, (anonymized) raw data should be shared, along with parameters and code to replicate the analysis, including proper significance testing. Prior to collecting human evaluation data, researchers should create an analysis plan and run power analyses to determine an appropriate sample size (likely requiring more workers and items than is currently typical in NLP).

7 Conclusion

Recent progress in NLP has been extraordinarily rapid, sometimes at the cost of experimental rigor. In this paper, we have presented evidence that underpowered experiments are widespread in NLP. For comparisons based on small samples, there is little reason to think that such an evaluation *could* reliably provide evidence of a significant improvement, and good reason to believe that improvements found to be significant will exaggerate or reverse the true effect. Going forward, a combination of larger test sets, simple power analyses, and wider sharing of code, data, and experimental details will help to build the foundation for a higher standard of experimental methodology in NLP.

Acknowledgments

Toyota Research Institute ("TRI") provided funds to assist the authors with their research but this article solely reflects the opinions and conclusions of its authors and not TRI or any other Toyota entity. Thanks to Sam Bowman, Amanpreet Singh, Kevin Clark, Naman Goyal, and Colin Raffel for providing data from submissions to the GLUE leaderboard, as well as Taylor Berg-Kirkpatrick, Sumanth Dathathri, Ari Holtzman, Hannah Rashkin, and Nikita Srivatsan for providing raw human evaluation data, not all of which made it into the paper.

[13] These simulations require estimates for 7 parameters: the baseline, the effect size, variance by worker, variance by worker as a function of model, variance by item, variance by item as a function of model, and residual variance.

References

Hua Ai, Antoine Raux, Dan Bohus, Maxine Eskenazi, and Diane Litman. 2007. Comparing spoken dialog corpora collected with recruited subjects versus real users. In *Proceedings SIGdial*.

Frank J. Anscombe. 1954. Fixed-sample-size analysis of sequential observations. *Biometrics*, 10:89–100.

Matthias G. Arend and Thomas Schäfer. 2019. Statistical power in two-level models: A tutorial based on Monte Carlo simulation. *Psychological methods*, 24(1):1–19.

Erfan Sadeqi Azer, Daniel Khashabi, Ashish Sabharwal, and Dan Roth. 2020. Not all claims are created equal: Choosing the right statistical approach to assess hypotheses. In *Proceedings of ACL*.

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second PASCAL recognising textual entailment challenge. In *Proceedings of the second PASCAL challenges workshop on recognising textual entailment*.

Dale J. Barr, Roger Levy, Christoph Scheepers, and Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278.

Colin B. Begg and Madhuchhanda Mazumdar. 1994. Operating characteristics of a rank correlation test for publication bias. *Biometrics*, 50(4):1088–1101.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The fifth PASCAL recognizing textual entailment challenge. In *Proceedings of TAC*.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in NLP. In *Proceedings of EMNLP*.

Katherine S. Button, John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson, and Marcus R. Munafò. 2013. Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5):365–376.

Boxing Chen and Colin Cherry. 2014. A systematic comparison of smoothing techniques for sentence-level BLEU. In *Proceedings of WMT*.

Kevin Clark, Minh-Thang Luong, Urvashi Khandelwal, Christopher D. Manning, and Quoc V. Le. 2019a. BAM! Born-again multi-task networks for natural language understanding. In *Proceedings of ACL*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2019b. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *Proceedings of ICLR*.

Jacob Cohen. 1962. The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, 65(3):145–153.

John E. Connett, Judith A. Smith, and Richard B. McHugh. 1987. Sample size and power for pair-matched case-control studies. *Statistics in Medicine*, 6(1):53–59.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. In *Proceedings of the Machine Learning Challenges Workshop*.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *Proceedings of ICLR*.

Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*.

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

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. In *Proceedings of EMNLP*.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of ACL*.

Stephen W. Duffy. 1984. Asymptotic and exact power for the McNemar test and its analogue with R controls per case. *Biometrics*, 40:1005–1015.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of EMNLP*.

Morten W. Fagerland, Stian Lydersen, and Petter Laake. 2013. The McNemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional. *BMC Medical Research Methodology*, 13.

Cristina Garbacea, Samuel Carton, Shiyan Yan, and Qiaozhu Mei. 2019. Judge the judges: A large-scale evaluation study of neural language models for online review generation. In *Proceedings of EMNLP*.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of ICML*.

Andrew Gelman. 2019. Don't calculate post-hoc power using observed estimate of effect size. *Annals of Surgery*, 269(1):e9–e10.

Andrew Gelman and John Carlin. 2014. Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, 9(6):641–651.

Andrew Gelman and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Andrew Gelman and Eric Loken. 2013. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*.

Yvette Graham, Nitika Mathur, and Timothy Baldwin. 2014. Randomized significance tests in machine translation. In *Proceedings of WMT*.

Tatsunori B Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. In *Proceedings of NAACL*.

John M. Hoenig and Dennis M. Heisey. 2001. The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician*, 55(1):19–24.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *Proceedings of ICLR*.

John P. A. Ioannidis. 2019. What have we (not) learnt from millions of scientific papers with P values? *The American Statistician*, 73(sup1):20–25.

John P. A. Ioannidis, T. D. Stanley, and Hristos Doucouliagos. 2017. The power of bias in economics research. *The Economic Journal*, 127(605):F236–F265.

Shankar Iyer, Nikhil Dandekar, and Kornél Csernai. 2017. First Quora dataset release: Question pairs.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP*.

Helena C. Kraemer and Christine Blasey. 2015. *How Many Subjects?: Statistical Power Analysis in Research*. SAGE.

Peter A Lachenbruch. 1992. On the sample size for studies based upon McNemar's test. *Statistics in Medicine*, 11(11):1521–1525.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *Proceedings of ICLR*.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of INLG*.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd schema challenge. In *Proceedings of KR*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. ROBERTA: A robustly optimized BERT pretraining approach. *Computing Research Repository*, arXiv:1907.11692.

R. Thomas McCoy, Junghyun Min, and Tal Linzen. 2019. BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance. *Computing Research Repository*, arXiv:1911.02969.

Blakeley B. McShane, David Gal, Andrew Gelman, Christian Robert, and Jennifer L. Tackett. 2019. Abandon statistical significance. *The American Statistician*, 73(sup1):235–245.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of WMT*.

Daniel J. O'Keefe. 2007. Brief report: Post hoc power, observed power, a priori power, retrospective power, prospective power, achieved power: Sorting out appropriate uses of statistical power analyses. *Communication Methods and Measures*, 1(4):291–299.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. FAIRSEQ: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL*.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *Proceedings of WMT*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of ACL*.

Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence encoders on STILTs: Supplementary training on intermediate labeled-data tasks. *Computing Research Repository*, arXiv:1811.01088.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of WMT*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *Computing Research Repository*, arXiv:1910.10683.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of ACL*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of EMNLP*.

Stefan Riezler and John T. Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*.

Jeffrey D. Scargle. 1999. Publication bias: The "file-drawer" problem in scientific inference. *arXiv*, arXiv:physics/9909033.

James J. Schlesselman. 1982. *Case-control studies: Design, conduct, analysis*. Oxford University Press.

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2019. Green AI. *Computing Research Repository*, arXiv:1907.10597.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of EMNLP*.

Anders Søgaard, Anders Johannsen, Barbara Plank, Dirk Hovy, and Hector Martínez Alonso. 2014. What's in a p-value in NLP? In *Proceedings CoNLL*.

Samy Suissa and Jonathan J. Shuster. 1991. The 2 x 2 matched-pairs trial: Exact unconditional design and analysis. *Biometrics*, 47(2):361–372.

Denes Szucs and John P. A. Ioannidis. 2017. Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS biology*, 15(3).

Eric-Jan Wagenmakers. 2007. A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, 14:779–804.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the Workshop on BlackboxNLP*.

Jacob Westfall, David A. Kenny, and Charles M. Judd. 2014. Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Pyschology: General*, 143(5):2020–2045.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of NAACL*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Proceedings of NeurIPS*.

Georgios N. Yannakakis and Héctor P. Martínez. 2015. Ratings are overrated! *Frontiers in ICT*, 2.