Researcher-Centered Design of Statistics: Why Bayesian Statistics Better Fit the Culture and Incentives of CHI

ABSTRACT

A core tradition of HCI lies in the experimental evaluation of the effects of techniques and interfaces to determine if they are useful for achieving their purpose. As a community we are inconsistent in publishing replication studies or statistical meta-analyses that more robustly demonstrate studied effects. Individual analyses tend to stand alone, and the quantitative knowledge from those studies does not accrue to more precise estimates via meta-analysis. We treat this as a user-centered design problem: the failure to accrue quantitative knowledge is not the users' (i.e. researchers') failure, but a failure to consider those users' needs when designing statistical practice. We use simulations to compare hypothetical publication worlds following existing frequentist against Bayesian practice. We show that Bayesian analysis allows us to estimate more precise effects with each new study, which supports knowledge accrual without traditional meta-analyses. Bayesian practices also allow more principled conclusions from small-*n* studies of novel techniques. These advantages make Bayesian practices a better fit for the culture and incentives of the field. Instead of admonishing ourselves to spend resources running larger studies, we propose using tools that more appropriately analyze small studies and encourage effective knowledge accrual from one study to the next without meta-analysis. Bayesian methods can be adopted from the bottom up without the need for top-down incentives for replication or meta-analysis, which suggests these techniques are a more user-(i.e. researcher-) centered approach to statistical analysis.

Author Keywords

Replication; meta-analysis; Bayesian statistics; small studies.

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous;

INTRODUCTION

A core focus of the CHI community is on the development of novel ideas and technology artifacts. The focus on novelty is valuable as it establishes a tradition that challenges assumptions about the design of technical systems and often results in insights that translate into more useful, usable, and enjoyable technologies. We, as a community, also value accurately understanding and precisely characterizing the effects of technology, which can be at odds with the first goal. Our community has limited resources to meet the latter goal as we have historically devoted more resources towards novelty and innovation over accurate and precise estimates of the utility of our systems.

Greater emphasis on novelty has led to concerns about the reliability of knowledge accrued in our field. To support quantitative knowledge accrual, we have adopted particular statistical tools, such as frequentist null hypothesis significant testing[1], and quantitative standards, such as $p < .05$, that define what constitutes sufficient evidence for researchers to support their claims. It is well known in the statistics community that results from individual studies—especially with the small sample sizes typical in our community—regularly fail to reliably estimate true effects [10]. To gain more reliable estimates of effects, multiple studies can be aggregated using frequentist meta-analytic techniques, which combine the results from multiple papers to obtain more precise effect size estimates (e.g., the difference between two conditions). However, our community rarely conducts meta-analyses. Paralleling so-called *replication crises* in psychology and medicine, movements such as RepliCHI [19] have called for an increased focus on replication and meta-analysis to effectively accrue quantitative knowledge about the utility and generalizability our technologies. Others emphasize deeper changes to statistical practice, such as reducing the focus on *p*-values in favor of effect size estimates and confidence intervals (the "New Statistics" [1]; or at CHI, Kaptein & Robertson [11]), or the abandonment of frequentist null hypothesis significance testing (NHST)[2] statistics altogether for Bayesian analyses.

[1] We will use frequentist and NHST interchangeably.

[2] While there exist Bayesian formulations of NHST based on Bayes factors, we believe they share some problems with frequentist NHST, such as a focus on binary testing

All of these suggestions, apart from a Bayesian approach, logically extend NHST, the dominant statistical approach used in CHI and related fields such as psychology. NHST is a statistical approach that turns research into a binary question: can we reject or fail to reject a *null hypothesis (i.e.,* that there is no effect)? A common use of NHST in CHI is to compare a novel system to a control system and, if a *p*-value is below the customary $p<.05$ target, then the new system is deemed better, as this *p*-value suggests that the observed difference is not likely due to chance. In other words, we reject the null hypothesis of no difference.

However, getting a p-value less than .05 can still happen even when there is no true meaningful difference. Replication and meta-analysis allow us to reduce this error in NHST (for example, the probability of falsely rejecting the null hypothesis) by combining the results of many studies of the same phenomenon. However, this requires at least one additional study, the meta-analysis, which necessitates new top-down incentives for conducting and publishing meta-analysis in CHI. By contrast, Bayesian analysis incorporates prior knowledge from other studies of the same and similar phenomena into a paper's quantitative analysis. A series of papers analyzing novel contributions can plausibly accrue knowledge and bypass the need for publishing separate meta-analyses. This allows increased precision of knowledge from the bottom up, within the existing publishing incentives of the field.

We consider the choice of statistical tools to be a user-centered design problem, with researchers as the users. Insisting that we *should* conduct meta-analysis amounts to blaming the users instead of the tools. Instead, we propose changing the tools—from NHST statistics to Bayesian statistics—in order to make quantitative accrual of knowledge easy (and even preferable!) within the existing publishing incentives of CHI. It is not researchers, but the statistical tools they have been given, that currently prevents this. Bayesian statistics are more user-centered statistics.

In the rest of this paper, we first give some background on replication and meta-analysis, and briefly compare Bayesian and NHST approaches to statistics. We then examine a subset of the HCI literature in the ACM digital library in order to assess the current state of meta-analyses in the field and establish that current incentives do not encourage meta-analyses, especially at the most prestigious venues. We then run several sequences of simulated experiments, representing hypothetical experiments run for separate publications, using a realistic effect size drawn from an existing meta-analysis. We then contrast two hypothetical publication worlds: one in which the simulated experiments were each analyzed in a traditional (NHST) manner (as would occur now), and one in which they were analyzed using Bayesian techniques. We demonstrate:

1. **The current state of quantitative knowledge accrual in HCI is poor**. Through an examination of publications in the ACM digital library, we demonstrate that little meta-analysis is conducted in the community.
2. Bayesian analysis **provides more precise estimates of previously-studied conditions in each successive study**. The NHST approach only increases the precision of effect sizes if the new study has a larger sample or when a meta-analysis is conducted. In contrast, the Bayesian approach uses prior knowledge to increase the precision of effect sizes for known conditions in each successive study, without requiring a meta-analysis (which is unlikely to be done at CHI).
3. Bayesian analysis **allows more precise comparison of novel conditions against known conditions**. By giving more precise effect size estimates of previously-studied conditions, Bayesian analysis increases the precision of estimated differences between existing and novel conditions.
4. Bayesian analysis **facilitates quantitative knowledge accrual within CHI's existing publishing incentives**. Unlike frequentist analysis, Bayesian analysis can accrue knowledge within individual studies without new top-down incentives for publishing meta-analyses, shifting knowledge accrual into original papers.
5. Bayesian analysis **draws more reasonable conclusions from small-*n* studies**. Bayesian analysis allows more principled estimates from small-sample studies of novel techniques by incorporating prior knowledge, and makes better use of prior knowledge so that researchers need not spend limited resources on larger studies to increase precision. This makes it particularly attractive to design and engineering researchers running small studies on novel technology
6. Bayesian analyses **shift the conversation from "Does it work?" to "How strong is the effect?", "How confident are we in this estimate?", and "Should we care?"** While NHST does incorporate effect size estimates and confidence intervals, ultimately, the use of a p-value (a fundamental feature of NHST) translates all questions into binary answers; can we reject or fail to reject the null hypothesis? Bayesian statistics emphasizes the likelihood of an effect given prior knowledge, thus shifting the conversation from "does it work?" to "how strong is the effect?" and "how confident are we?" NHST can technically answer these questions but under-emphasizes them. Bayesian statistics better emphasizes the ultimate questions of our work: should we care about these results enough as practitioners to adopt new designs or as researchers to study this more?

BACKGROUND AND MOTIVATION

In this section we first discuss the current state of meta-analysis and replication at CHI and how it is dictated by the

rather than precision of estimation and cost-benefit analysis; thus we do not consider them here.

community's publication incentives. We then introduce the basics of Bayesian analysis as compared to frequentist.

Interpretation of Bayesian versus frequentist statistics

Interpretations of frequentist statistics are a common source of errors amongst CHI researchers (and others). The focus on p-values/significance testing amounts to insisting that users learn how to interpret the cognitively demanding conceptual double-negative of a *p* value, instead of interpreting results as evidence for a hypothesis—a valid interpretation within a Bayesian framework [13]. This interpretation problem has resulted in simplifying the results of NHST tests into the belief that the p-value answers the question, "Does it work?". Unfortunately, this is an inaccurate interpretation and lies at the heart of the argument to shift interpretation towards effect sizes and confidence intervals [1] (though 95% confidence intervals are the inverse of $p<.05$, thus still succumbing to the interpretation problem [8]). Effect size estimates and confidence in those estimates is the essential information sought, but NHST logic unintentionally relegates this information as secondary to the p-value.

By contrast, Bayesian analysis provides formal approaches to quantifying our existing beliefs (for example, as a probability distribution over the expected difference in the means of some variable between two conditions), and then updating those beliefs based on new experimental evidence. This gives results expressed as probabilistic evidence for or against hypotheses, emphasizing effect size estimates and confidence in the estimates. This information supports the decisions researchers and practitioners are making with the data: should I incorporate this technology into my practice or, if confidence is low, are the results promising enough to continue to study it? Bayesian statistics are more user-centered as they emphasize the information needed to support the actual decisions of researchers without the need to interpret double-negative logic.

Replication and meta-analysis in CHI

The statistical tools researchers customarily use in CHI do not help them effectively accrue knowledge from one study to the next, even when the variations in design of novel systems are informed by previous work. The classic strategy for knowledge accrual of a series of NHST studies is literature reviews conducted often in the related work section of a CHI paper that implicitly use the *vote-counting* method of knowledge accumulation [7]. In this method, the number of significant and non-significant findings are counted up to infer if an effect is true or not (e.g., three studies found a significant effect, four did not, therefore this strategy is likely not effective). There are many problems with this approach, particularly when a field utilizes small samples to estimate statistical significance as many of these significant differences are likely due to chance [10]).

A step towards better knowledge accrual is via the use of meta-analyses, where the focus is not on the statistical significance of any single study, but instead on combining the results from many studies to estimate the *effect size* (e.g., this system designed for encouraging exercise results in 1,000 more steps per day compared to control) and the confidence in that effect (i.e., that 1,000 step increase could feasibly be as low as 100 steps or as high as 1,900 steps). This strategy relies somewhat on increasing the incentives for replication in the literature, an approach currently advanced by RepliCHI [19]. While encouraging more standalone replication studies and meta-analyses is useful for knowledge accrual, we argue that it has difficulty fitting into CHI culture and the incentives for publishing novel findings. As we will show, few meta-analyses are currently conducted in the community, supporting our intuition.

Within a Bayesian approach, prior beliefs can be derived from previous work, allowing knowledge accrual from study to study without requiring a separate meta-analysis. To derive priors in CHI, we can capitalize on the fact that partial replication is common to the field in the form of the comparison of a new technique against the state-of-the-art. As we will show, incorporating prior quantitative results into new analyses using a Bayesian framework is straightforward in these cases, allowing us to accrue quantitative knowledge without the need for top-down incentives for meta-analysis. We will also discuss how to use prior work to set prior expectations on the size of an effect even when not conducting a partial replication. In contrast to traditional meta-analysis, Bayesian analysis allows the effect sizes in successive studies to be estimated more precisely in each study. This fits well into the publishing incentives for CHI: knowledge accrues with each individual, novel study (easily published at CHI), making it unnecessary to publish standalone meta-analyses (less publishable at CHI).

HCI RESEARCHERS CONDUCT FEW META-ANALYSES

To assess the current state of quantitative knowledge aggregation in HCI, we conducted a review of meta-analyses accessible through the ACM Digital Library, as many of the most prominent HCI publication venues are archived there (e.g. CHI, CSCW, UIST, UbiComp, TOCHI). We searched for the terms *meta-analysis*, *meta-analyses*, *metaanalysis*, or *metaanalyses* in the abstract or title fields on Aug 17 2015, yielding 509 unique results. We examined abstracts and eliminated 151 domain-specific statistical methods and techniques, mostly in biology and machine learning. We examined the full-text of the remaining papers. We found 40 dissertations, which we discarded since their results may have been published in other venues. We found 56 papers with quantitative meta-analyses, defined as modeling effect sizes or using traditional meta-analysis based on the results of multiple studies found from a literature search with inclusion criteria. Only 3 were published at the venues above [9,18,21] This low number prompted us to search the DL full text for "meta-analysis" for the top venues, yielding 159 results. The top 3 results were the meta-analyses we had already found, and we did not find any others after reviewing the abstracts (and full text as needed) from this additional search. Most meta-analyses were in other journals and communities, from management information sys-

tems and HICSS to specialized venues (e.g. ICMI '06: Proceedings of the 8th international conference on Multimodal interfaces).

Our search suggests that meta-analyses are not being rewarded by the current publishing incentives of the community. The CHI conference is regarded by many as the top publication venue for work in human-computer interaction. Given the paucity of meta-analysis at CHI, it seems clear that the incentives do not currently exist for such work to be published there. However, within the frequentist paradigm, a meta-analysis is the gold standard for quantitative knowledge accrual, representing the best estimates we can make. We must either build new incentives for meta-analysis, or find another way to accrue knowledge within the existing incentives. That other way is Bayesian analysis.

CONTRASTING FREQUENTIST AND BAYESIAN KNOWLEDGE ACCRUAL USING SIMULATED EXPERIMENTS

By way of explaining the differences between frequentist meta-analysis and a Bayesian incremental approach to knowledge accrual, in this section we provide an example of these approaches applied in two different hypothetical worlds. Specifically, we examine a series of 4 simulated, hypothetical experiments on the effects of progress indicators on completion rates of online surveys.

Domain: Varying progress indicators in online surveys

We chose this domain because it will be familiar to the CHI audience (as many researchers in our field make use of online surveys), and because a meta-analysis has previously been conducted in this domain by Villar *et al.*[17], providing us with realistic effect sizes to use in our simulations.

That meta-analysis looked at experiments comparing the effects of different types of *progress indicators* on survey completion rates. A progress indicator is any type of textual or graphical display communicating how much of the survey has been completed so far ("10%", a graphical progress bar, etc). Progress indicators can be distinguished by the relationship between the true progress and the displayed progress. A *constant* indicator communicates the true progress. Progress in a *fast-to-slow* indicator starts fast, telling the participant they have made more progress than they actually have near the beginning of the survey, then slows down later. By contrast, a *slow-to-fast* indicator starts slow, then speeds up near the end of the survey.

In their meta-analysis, Villar *et al.* [17], found that using a *slow-to-fast* progress indicator decreased the probability that a person would complete the survey. They found an effect size (as a log odds ratio[3]) of ~-0.45.[4] A log odds ratio of -0.45 means that in a survey that would otherwise have a completion rate of 50%, we would expect the same survey with a *slow-to-fast* progress indicator to have a completion rate of ~39%.

Simulation Method

To compare Bayesian and frequentist approaches, we will simulate 100 hypothetical "worlds" in which we know the true effect of different progress bar types on completion rates, and then run the same series of 4 experiments in each world. Each experiment could represent an experiment run by different authors, thus representing the error that would be present from a single study, the current primary mechanism for estimating effect sizes. We will conduct analyses on each world as if 1) all authors take a frequentist approach or 2) all authors take a Bayesian approach.

For the purposes of our simulations, we will consider the true effect of a *slow-to-fast* progress bar on the log-odds of the completion of a survey to be -0.45, as suggested by the meta-analysis of Villar *et al.* [17]. We will also surmise a similarly-sized effect of *fast-to-slow* progress indicators, in the opposite direction, of 0.45.[5] This approach supports examination of knowledge accrual in both worlds and allows comparison to the known fact (because we define it in the simulation) that there is an effect size of -0.45 (*slow-to-fast*) and 0.45 (*fast-to-slow*) compared to no indicator.

In each world, we simulate the results of 4 experiments:

- **Experiments 1-3** have between-subjects designs comparing a *fast-to-slow* progress indicator against a control condition of no indicator.
- **Experiment 4** also compares a *fast-to-slow* progress indicator against a control condition, but adds an additional *slow-to-fast* indicator. We can think of this experiment as representing one of the common ways that partial replication happens in the CHI community: through comparison to previous state-of-the-art results. Perhaps some authors, having seen the success of *fast-to-slow* indicators, wished to know how the opposite type of indicator might perform (or perhaps conducted this experiment as part of work to establish a more complete theory explaining *why* we see these particular results).

---

[3] A log odds ratio of 0 indicates no difference between conditions. The log odds ratio is the log of the ratio of the odds of someone completing the survey in one condition compared to another condition. It is regularly used in comparing probabilities between two conditions because (unlike, say, differences of proportions), it is unbounded, which simplifies analysis. It is related to logistic regression in that coefficients of a logisitic regression can be interpreted as log odds ratios.

[4] While their results use probability of drop-out, we use probability of completion.

[5] It is worth noting that this effect may be larger than the true effect in the real world, since Villar et al. [17] found the *fast-to-slow* effect is likely closer to .3; but for our purposes it remains a realistic effect size based on the *slow-to-fast* results, and simplifies the example by mirroring *slow-to-fast*.

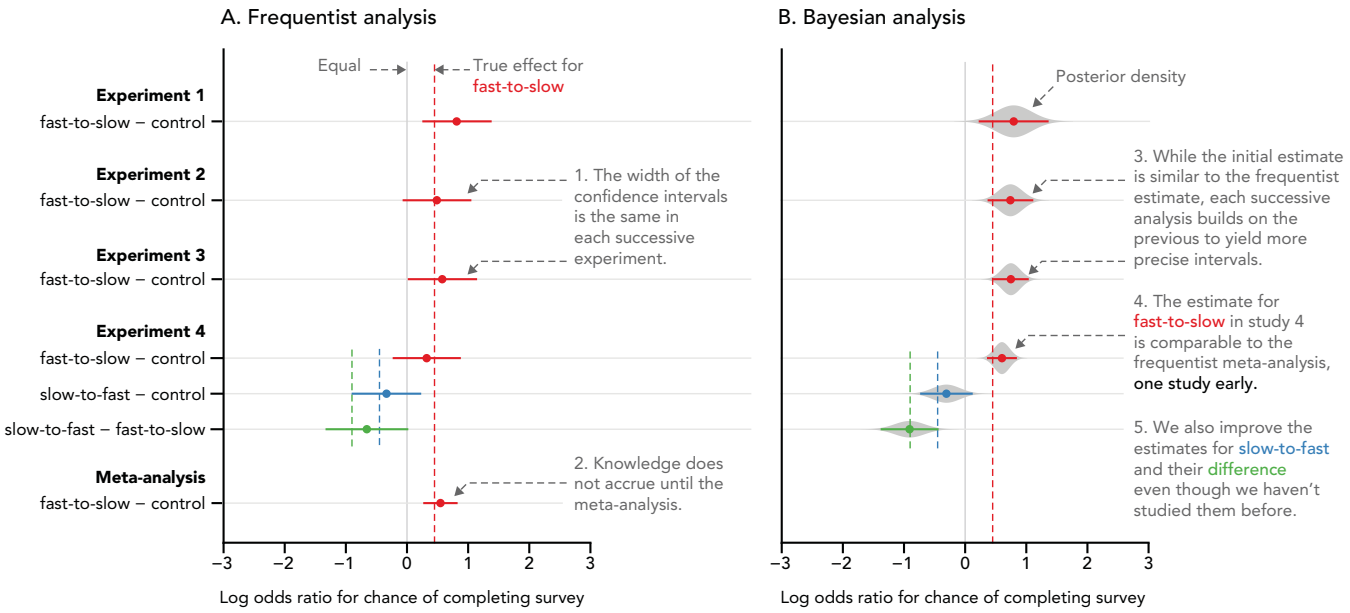

Figure 1. Forest plots of effects from the frequentist and Bayesian analyses applied to one of our simulated worlds with 100 participants per condition.

For simplicity of exposition, we assume the same experimental design in each case: a between-subjects design with 100 participants per condition (thus, 200 participants in experiments 1-3 and 300 in experiment 4). This is similar to the number of participants in the studies in Villar *et al.*'s meta-analysis, and is a reasonable number to expect to respond to an online survey. The between-subjects design is necessary primarily because it is difficult to ask someone to take the same survey twice and observe their drop-out rates.

Frequentist analysis

In the frequentist analysis, we conduct a logistic regression in each experiment to model the probability of *completion* based on the *progress indicator condition*: *control* (no indicator), *fast-to-slow*, and *slow-to-fast* (experiment 4 only). In addition, after all four experiments are analyzed, we conduct a meta-analysis on the log-odds ratios for the effect of the *fast-to-slow* progress indicator, as in Villar *et al.* [17]. This yields a final, more precise estimate of the effect of that indicator based on the preceding four experiments.

Bayesian analysis

In the Bayesian analysis of each world, we also conduct a logistic regression in each experiment to model the probability of *completion* based on *progress indicator*. However, we do not conduct a final meta-analysis. Instead, starting with experiment 2, we build upon the previous results by using the posterior distribution (i.e., estimates from the previous study) of the estimated effect of the *fast-to-slow* progress indicator in experiment *i* as the prior for that effect in experiment *i + 1*. We assume this could happen, for example, if the author of experiment *i + 1* had read the previous paper, and therefore was able to use the posterior estimate from that paper in their analysis (previous study's posterior becomes the next study's prior). This results in incremental knowledge accumulation without a meta-analysis.

In experiment 4, we must also place a prior on the new *slow-to-fast* indicator. We use a Cauchy[6] distribution centered at 0 (no effect) with a scale equal to the furthest point in the 95% credibility interval[7] of the estimated effect of *fast-to-slow* in experiment 3. This prior is weakly-informed: it expresses a belief that fast-to-slow might reasonably have about twice the effect (positively or negatively) that *slow-to-fast* does compared to control condition. While it is beyond the scope of this paper to discuss the various strategies for setting priors, this is a core topic in any book on Bayesian analysis (see e.g. [14]).

Results

In a single world

Before contrasting the Bayesian and frequentist results across all simulated worlds, we will first walk through the results of one simulation. The results of the frequentist analysis are shown above in Figure 1A and the results of the Bayesian analysis in Figure 1B. Each figure shows a

[6] The Cauchy distribution is similar to the Normal distribution, but with fatter tails. Gelman recommends it for use as a weakly-informed prior because the fatter tails express less certainty in the location of the effect [3].

[7] Roughly, a credibility interval is the Bayesian analog to a confidence interval. Unlike a confidence interval, however, a proper credibility interval *is* an expression of the probability of the location of a parameter (the confidence interval is not, despite its common misinterpretation [8]; again a reason frequentist statistics are not user-friendly).

forest plot of results, with 95% CIs (*confidence intervals* in the frequentist case; *credibility intervals* in the Bayesian case). In the frequentist case, a 95% confidence interval that does not overlap 0 is equivalent to a p value of less than 0.05, suggesting that the null hypothesis (i.e., no effect) can be rejected with 95% confidence. The dashed vertical lines indicate the true effect sizes from which the data was simulated.

In the frequentist analysis, we have a promising first result in experiment 1. This is followed by two borderline results in experiments 2 and 3. Looking strictly at p values, experiment 4 fails to replicate the result of experiment 1, though it does find some evidence of a difference between *fast-to-slow* and *slow-to-fast* progress indicators. Finally, the meta-analysis is able to combine the previous estimates into a more precise and accurate estimate of the true effect --- assuming it is conducted and published.

Note that, because all of these experiments are run using the same number of participants, the confidence intervals are all approximately the same width; the only ways to increase our precision (i.e., decrease CI width) in the frequentist world are to increase the power of our experiment/analysis (for example by increasing our sample size, using a within-subjects design, or including covariates that explain some of the variation in the response) or by conducting meta-analysis. This limitation is not particularly helpful to the authors of experiments 1-4, since they may not have the resources to recruit more participants.

In addition, the small variations in intervals from experiments 1-3 represent vastly different conclusions if we reduce the results to null hypothesis tests: experiments 1 and 3 reject the null ($p < 0.05$); experiment 2 does not. This highlights the problem with reducing estimation to a binary choice ("effect" or "no effect"): these estimates are all similar, but the decision to reject (or not) the null hypothesis hinges on whether the 95% confidence interval overlaps 0 in that particular study, thus resulting in false conclusions as per the vote-counting method discussed earlier (current CHI standard practice).

In the Bayesian analysis, the result of the first experiment is virtually identical to the frequentist world (we used a weakly-informed Cauchy(0, 2.5) prior for logistic regression parameters recommended by Gelman *et al.* [3]). However, in contrast to the frequentist world, in each subsequent experiment our estimate of the effect size becomes more precise. The authors of experiments 2 and 3 make a stronger contribution to the field by building on the results of prior work, rather than borderline failed replications. In experiment 4, the estimated effect of the fast-to-slow indicator is similar to that of the frequentist meta-analysis, *one publication early*. **Bayesian analysis helps us learn faster and with fewer studies**.

Besides the benefit of getting quantitative knowledge accrual into the literature without requiring publication of

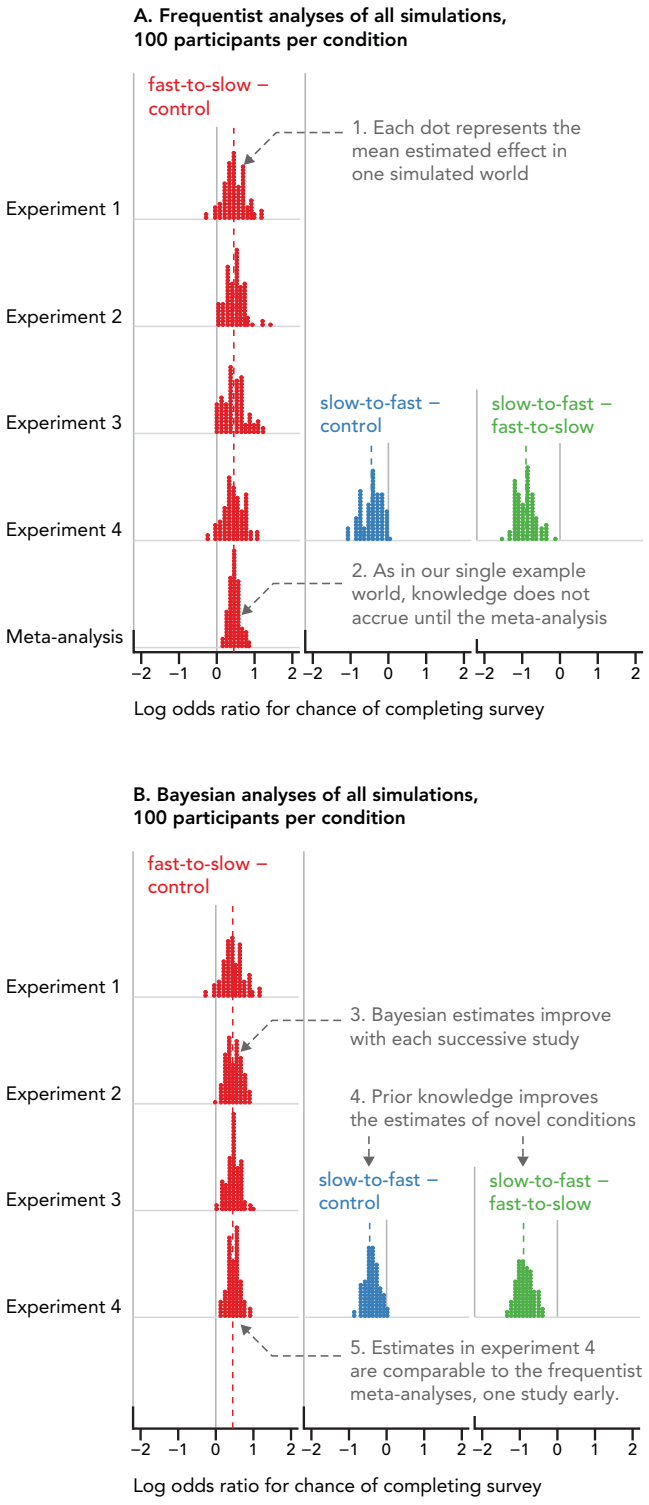

Figure 2. Results of the Frequentist (A) and Bayesian (B) analyses of all simulated worlds, n=100 per condition.

meta-analysis, this also has additional benefits for the authors of experiment 4: note that, even though they are testing a new technique that they don't have strong priors for (the *slow-to-fast* indicator), the strong prior knowledge of the effect of the *fast-to-slow* indicator helps them estimate

the effect of the novel technique more precisely. This is because the more precise estimate of *fast-to-slow* also helps makes the estimate of the difference between *fast-to-slow* and *slow-to-fast* a little more precise. In other words, more precise estimates of techniques we've seen lead to more precise estimates of comparisons to new techniques, which even makes estimates of those new techniques a little more precise. **Bayesian analysis helps us apply old knowledge to novel questions.**

In many worlds

We now step up to consider the effects of the two analysis approaches in all 100 simulated worlds. Each point in Figure 2A represents the mean estimated effect from the frequentist analysis in one of the simulated worlds. Figure 2B shows the mean estimated effects from the Bayesian analyses.

We can see that the pattern observed in our single example world holds true across simulations: the estimated effect becomes more precise with each experiment in the Bayesian analysis, and the final estimate for *fast-to-slow* resembles the frequentist meta-analysis, one study early. In addition, the estimates for *slow-to-fast* are more precise in the Bayesian analysis of experiment 4 due to the use of prior knowledge, even though we have never seen that condition before. This is reflected in the root-mean-squared error of those estimates compared to their true effects in experiment 4:

	Frequentist	Bayesian
fast-to-slow − control	0.27	0.17
slow-to-fast − control	0.27	0.20
slow-to-fast – fast-to-slow	0.26	0.22

Table 1. Root mean-squared error of estimates in experiment 4 with 100 participants per condition.

Note that the estimated effects in the frequentist analysis all have approximately the same error, reflective of the power of the experiment. The Bayesian approach gives us estimates with less error by building knowledge as we go.

BAYESIAN ANALYSIS OF SMALL SAMPLES

Due to limited resources and an emphasis on novelty, HCI studies are often conducted with fewer participants than a traditional power analysis would suggest is prudent [4,11]. With a frequentist analysis, this increases the probability of what Gelman calls a *magnitude error* [2]: because the confidence intervals are so wide, the only effects that reach significance are those that overestimate the effect size.

That said, we believe that there are many reasons why small-*n* studies are conducted in HCI, including limited resources and the emphasis on novelty discussed earlier as fundamental for contributions to the field. Thus, we ask: can Bayesian analysis help make better use of our limited

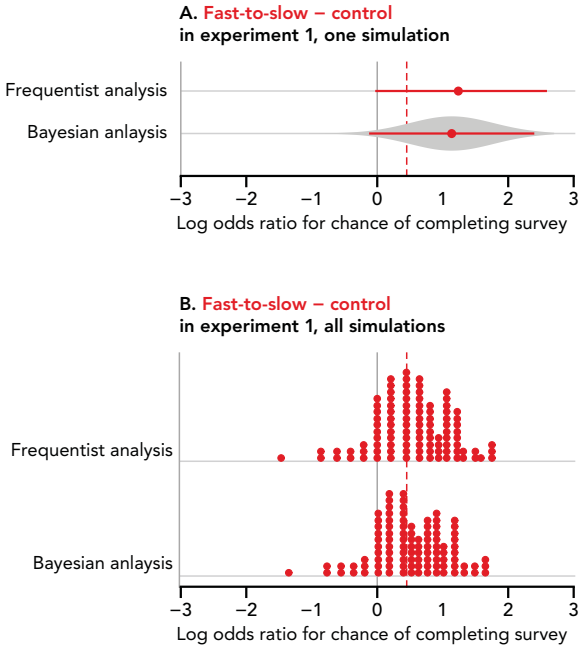

Figure 3. The effects of shrinkage in experiment 1 with 20 participants. A. From a single simulation, showing the shrinkage of a single estimate in the Bayesian analysis. B. The general shrinkage of estimates towards 0 from all simulations of experiment 1 with 20 participants.

resources? Can we do better than simply admonishing researchers to recruit more participants? To assess this, we repeated our simulations with 20 participants per condition instead of 100.

Results

At only 20 participants per condition, the precision of our estimates (width of the confidence interval) guarantees that any significant results of the frequentist analysis will hugely over-estimate the size of the effect. However, even though we haven't (in our hypothetical world) studied this particular effect before, we do have some prior knowledge about what constitutes small, medium, and large effects in studies of human behavior. In the Bayesian analysis, this knowledge is encoded in the prior we set on the effect; here, we adopt a weakly-informed prior from Gelman [3].

Bayesian analysis effectively weighs how strong our prior knowledge is against how much evidence we have: while our prior had only a small effect on estimates in the 200-participant experiments (where we had enough evidence to easily shift a diffuse prior), with only 20 participants our prior has more influence. In the frequentist world we might intuitively dismiss overly large effect sizes in small studies as unreasonable; in the Bayesian world we can consistently and quantitatively apply this intuition by encoding it in a prior and using the prior to shift unreasonably large effects towards zero. This is called **shrinkage**.

We can see the effects of shrinkage by comparing a large estimate from experiment 1 with its corresponding Bayesi-

an estimate (Figure 3). The Bayesian approach shrinks the unreasonably large estimate a little bit towards 0, reflecting our skepticism. The resulting posterior is still quite diffuse: we haven't learned all that much from the small study. But what we have learned is reasonable in proportion to what we knew before and how much evidence we have. This posterior still advances the knowledge of the field such that subsequent studies will be more precise --- even if it doesn't reach a frequentist notion of significance (note that to reach significance with samples of 20, the effect would need to be nearly 3 times the actual effect!).

In many worlds

We can see the effects of shrinkage on the first experiment when we look at all of the 20-participant simulations (Figure 4): the most extreme estimates in experiment 1 are moved slightly towards 0. This has the effect of reducing the overall error in experiment 1 by discounting unreasonably large estimates that occur due to chance:

	Frequentist	Bayesian
fast-to-slow − control	0.61	0.56

Table 2. Root mean-squared error of estimate in experiment 1 with 20 participants per condition, reflecting the effect of Bayesian shrinkage on discounting unreasonably large effects in small-*n* studies.

We also see the same narrowing of precision in successive studies in the Bayesian world as we did with the 100-participant simulations. By the time we reach experiment 4, the difference in error is dramatic: the estimate for *fast-to-slow* has nearly half the error in the Bayesian world (.36 versus .66), and we again get better estimates of the novel condition, *slow-to-fast*:

	Frequentist	Bayesian
fast-to-slow − control	0.66	0.36
slow-to-fast − control	0.68	0.51
slow-to-fast – fast-to-slow	0.83	0.60

Table 3. Root mean-squared error of estimates in experiment 4 with 20 participants per condition.

DISCUSSION

In this section we discuss several implications of our suggested approach to statistics in CHI.

Bayesian analysis increases the value of small-*n* studies of novel work

The traditional solution to the problems associated with low-power studies (and one HCI researchers are often admonished to adopt) is to spend resources recruiting more participants. In other words, the frequentist solution to low power is *not to run low-powered studies*.

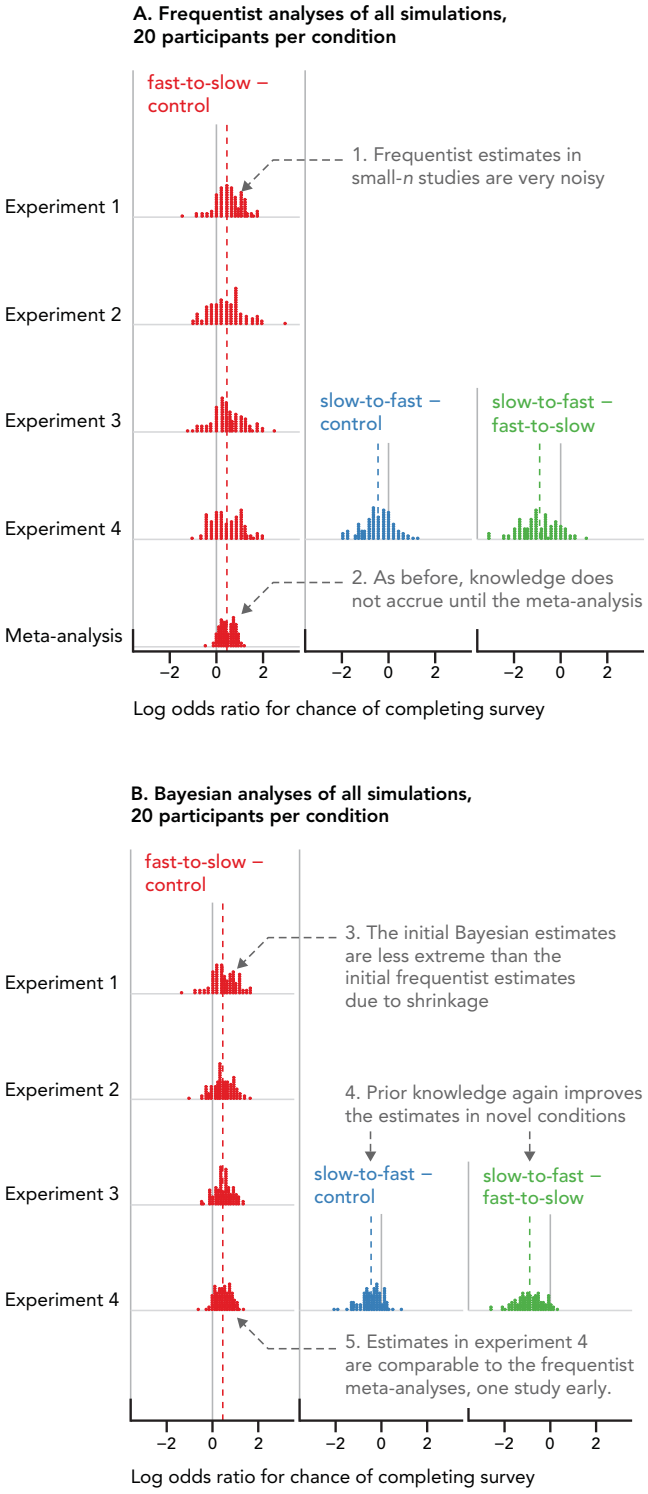

Figure 4. Results of the Frequentist (A) and Bayesian (B) analyses of all simulated worlds with 20 participants per condition.

However, researchers developing complex new systems or interaction techniques have the expertise, time, and resources for that type of work; spending their limited resources on running larger studies may be a poor allocation

of work across the research community. These researchers already (in our view, rightly) protest that they are asked to run *pro forma* evaluations when their primary contributions are in engineering or design (see e.g., Greenberg and Buxton [6]); telling them not only to run evaluations but to recruit more participants amounts to blaming the users.

Statistics can be a tool for communication and collaboration. Bayesian analysis better supports new kinds of collaborations that take better advantage of specialization. We see researchers that produce novel systems and interaction work as having a symbiotic relationship with others who have the resources and expertise for larger quantitative work (but perhaps not the expertise for novel engineering): the latter researchers might find a novel technique in the literature, adapt it to some domain based on users' needs, and evaluate it more extensively. In this context, the goal of small, early studies then becomes to demonstrate face validity of a technique and provide a rough first estimate of its effectiveness, not to find a (likely over-estimated in terms of magnitude) significant difference. For this, Bayesian analysis helps draw reasonable conclusions from small-*n* studies. It provides a more nuanced and accurate tool for evaluating contributions and combining them given the varying skills and resources of researchers.

Part of the goal of this paper is to release novel work in HCI from the chains of meaningless *p* values from small-*n* studies. We believe that small, early evaluations of novel work are still valuable, but that their output should be a probability distribution of expected effect size, whether or not it overlaps 0. "No effect" as determined via p-values should not be a barrier to publication of novel design work when we know that any effect that is found in a small study is likely overestimated or simply due to chance. Instead, the novel work should be (and already often is) judged on the merits of design and engineering, not a *pro forma* small-*n* evaluation. Bayesian analysis provides a richer conceptual understanding and role for these initial evaluations and helps to quantify information (i.e., effect sizes and confidence in those effect sizes) to support the questions implied by the community: should I incorporate this novel tool into my practice or, if not confident enough, is further research in this domain warranted?

Bayesian analysis fits into how statistical practice is shaped at CHI

The HCI community is large and multi-disciplinary; therefore, we believe that statistical practice at CHI is best shifted in a bottom-up fashion. For example, Wobbrock *et al.* [20] at CHI 2011 introduced a nonparametric analysis technique to the community --- the aligned rank transform (ART) --- applicable to various forms of data, including Likert scales. Since then, this approach has been widely adopted, and has been cited 148 times.[8] This adoption did not require new top-down incentives for improved analysis, but spread study-to-study and researcher-to-researcher.

Following the model of ART, we believe that Bayesian analysis can be adopted gradually in individual studies, sidestepping the difficulty of shifting an entire multi-faceted field from the top down. Statistical practice in scientific fields tends towards a model of mentorship and of drawing upon approaches found in prior work—e.g., as other papers begin adopting techniques like ART, readers of those papers will use similar techniques when conducting their own analyses in follow-up work. This is the candidate way to introduce Bayesian analysis: when readers see it used in a paper they wish to build upon, the analysis offers a direct way to do that: teaching by example. Such a paper also provides priors for the next researcher. In this manner such analyses can spread in the community, slowly building a body of work and a new standard of practice.

Bayesian analysis is accessible to practitioners

Even 15 years ago, Bayesian analysis was arguably impractical for most researchers due to a lack of tools and computational power. However, tools for building and running Bayesian models are now widespread, and have mature support in languages already used for data analysis, such as R and Python. These tools include modeling languages like JAGS [15] and Stan [16] (both with R packages, and Stan includes a Python interface), and Python-specific libraries like emcee (http://dan.iel.fm/emcee/current/) and PyMC (https://pymc-devs.github.io/pymc/). In addition, literature aimed at practicing researchers has made Bayesian modelling accessible: we particularly recommend Kruschke's *Doing Bayesian Data Analysis* [14] (which includes a table of common frequentist analyses and their Bayesian equivalents), as well as his proposed BEST test,[9] a robust Bayesian alternative to the *t*-test [12]. Other accessible articles have also been written about practical concerns in Bayesian analysis, including discussions of how to choose priors [3,5]. Still, statistical tools, whether frequentist or Bayesian, can exhibit high barriers to entry or silently fail, providing a poor interface. We believe there is a fruitful area of work in designing better tools and interfaces for statistical methods.

Practical impact of research through cost/benefit analysis

Finally, we wish to address another common thread of discussion in the CHI community, a perhaps more existential one: how can we have practical effects on real-world deployed systems? How can practitioners derive value from results at CHI? We believe that the language of statistical

[8] According to Google Scholar, accessed 2015-09-23: https://scholar.google.com/scholar?cites=16254127723353600671&as_sdt=5,48&sciodt=0,48&hl=en

[9] Somewhat glibly, BEST stands for *Bayesian estimation supersedes the t-test*

significance is not the language of practitioners or business; cost/benefit analysis is. The results of a Bayesian analysis can easily be incorporated into cost/benefit analysis: given the probability distribution of an estimated effect, we can simply apply a cost function to it.

For example, imagine a market research company that wishes to evaluate the cost/benefit of switching from an existing survey tool that does not have a *fast-to-slow* progress indicator to one that does. This would incur some costs for converting the survey into a new format. It would also have an estimated benefit in that the company could recruit fewer participants to reach a desired sample size, in proportion to the expected increase in completion rate. This company could take the probability distribution of estimated completion rate in both cases (*whether or not* the difference has passed the statistical significance filter) and use it to derive a probability distribution of expected cost in each case, and then decide a course of action to minimize cost. This simplifies the translation of research into real-world use, and gives a way to put practical effect sizes in context.

CONCLUSION

Bayesian analysis allows us to learn more quickly by building on previous results. It also fits more effectively into the publication incentives of CHI than approaches to improving knowledge accrual within the NHST framework, such as meta-analysis. At the same time, it is compatible with calls for more replication (RepliCHI), and allows us to make stronger claims about novel work through comparison to well-studied conditions. This, combined with a shift to an emphasis on probable effect sizes instead of simply "significant" differences between conditions, will help free design and engineering researchers from the shackles of meaningless p values in small-*n* studies, while also allowing the field to make better use of the results of such studies. In short, Bayesian statistics are user-centered statistics.

REFERENCES

1. Geoff Cumming. 2014. The new statistics: why and how. *Psychological science* 25, 1: 7–29. Retrieved July 9, 2014 from http://pss.sagepub.com/content/early/2013/11/07/0956797613504966.abstract
2. A. Gelman and J. Carlin. 2014. Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science* 9, 6: 641–651. Retrieved November 17, 2014 from http://pps.sagepub.com/content/9/6/641.short
3. Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. 2008. A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models. *The Annals of Applied Statistics* 2, 4: 1360–1383.
4. Andrew Gelman and David Weakliem. 2009. Of Beauty, Sex and Power. *American Scientist* 97, 4: 310–316.
5. Andrew Gelman. 2006. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* 1, 3: 515–533. http://doi.org/10.1214/06-BA117A
6. Saul Greenberg and Bill Buxton. 2008. Usability evaluation considered harmful (some of the time). *Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08*, ACM Press, 111. Retrieved September 25, 2015 from http://dl.acm.org/citation.cfm?id=1357054.1357074
7. Larry V. Hedges and Ingram Olkin. 1980. Vote-counting methods in research synthesis. *Psychological Bulletin* 88, 2: 359–369.
8. Rink Hoekstra, Richard D. Morey, Jeffrey N. Rouder, and Eric-Jan Wagenmakers. 2014. Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review* 21, 5: 1157–1164. Retrieved July 1, 2015 from http://www.ncbi.nlm.nih.gov/pubmed/24420726
9. Kasper Hornbæk and Effie Lai-Chong Law. 2007. Meta-analysis of correlations among usability measures. *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '07*: 617. http://doi.org/10.1145/1240624.1240722
10. John P A Ioannidis. 2005. Why most published research findings are false. *PLoS medicine* 2, 8: e124. Retrieved July 9, 2014 from http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124
11. Maurits Kaptein and Judy Robertson. 2012. Rethinking statistical analysis methods for CHI. *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*, ACM Press, 1105. Retrieved September 25, 2015 from http://dl.acm.org/citation.cfm?id=2207676.2208557
12. John K Kruschke. 2013. Bayesian estimation supersedes the t test. *Journal of experimental psychology. General* 142, 2: 573–603. Retrieved September 25, 2015 from http://www.ncbi.nlm.nih.gov/pubmed/22774788
13. John K. Kruschke. 2010. Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science* 1, 5: 658–676. http://doi.org/10.1002/wcs.72
14. John K. Kruschke. 2011. *Doing Bayesian Data Analysis*. Elsevier Inc.
15. Martyn Plummer. 2003. JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*. http://doi.org/10.1.1.13.3406
16. Stan Development Team. 2015. *Stan Modeling Language: User's Guide and Reference Manual*.
17. Ana Villar, Mario Callegaro, and Yongwei Yang. 2013. Where Am I ? A Meta-Analysis of Experiments

on the Effects of Progress Indicators for Web Surveys. *Social Science Computer Review* 00, 0: 1–19. http://doi.org/10.1177/0894439313497468

18. Suzanne Weisband and S Kiesler. 1996. Self Disclosure on Computer Forms : Meta-Analysis and Implications. *ACM Digital Library* CHI, 96: 3–10. http://doi.org/10.1145/238386.238387
19. Max Wilson, Wendy E. Mackay, Ed Chi, Michael Bernstein, and Dan Russell. RepliCHI - CHI should be replicating and validating results more: discuss. *CHI EA '11: Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems*. http://doi.org/10.1145/1979742.1979491
20. Jacob O Wobbrock, Leah Findlater, Darren Gergle, and James J Higgins. 2011. The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only ANOVA Procedures. *CHI '11*: 143–146.
21. Nick Yee, Nick Yee, Jeremy N Bailenson, Jeremy N Bailenson, Kathryn Rickertsen, and Kathryn Rickertsen. 2007. A meta-analysis of the impact of the inclusion and realism of human-like faces on user experiences in interfaces. *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '07*: 1. http://doi.org/10.1145/1240624.1240626