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

# Abstract

A core tradition of HCI lies in the experimental evaluation of the effects of techniques and interfaces. A well-acknowledged gap in the literature is the inconsistent publication of replication studies or statistical meta-analyses that more robustly demonstrate these studied effects. In other words, individual analyses tend to stand alone, and the quantitative knowledge from those studies does not accrue as it would in a meta-analysis. We treat this as a user-centered design problem, where the failure to accrue quantitative knowledge in the field is a not the users’ (i.e. researchers’) failure, but rather a failure to consider those users’ needs when designing statistical practice. Bayesian statistical methods, in contrast to currently-used frequentist methods, fit better into the needs of the community and facilitate quantitative knowledge accrual in the context of limited resources for evaluation and a focus on testing novel designs. We use simulations to compare hypothetical publication worlds following existing (frequentist) practice against Bayesian practice. We show that Bayesian analysis allows us to use prior knowledge to estimate more precise effects with each new study (and even incentivizes doing so), accruing knowledge without traditional meta-analyses and within the existing publication incentives of the field. We also demonstrate that the use of Bayesian statistics even allows us to draw more principled conclusions from small-*n* studies of novel techniques, a common practice at CHI that is hard to defend from a frequentist point of view. Instead of admonishing designers and engineers to spend resources running larger studies, we propose giving them the tools to more appropriately analyze small studies. We believe that Bayesian methods can be adopted from the bottom up at CHI, following the existing mentorship tradition in statistical practice, without the need for new top-down incentives for replication or meta-analysis. Ultimately, these techniques represent a more user- (i.e. researcher-) centered approach to statistical analysis.

# Introduction

A core focus of the CHI community is on the development of novel ideas and technology artifacts. The focus on novelty and innovation is valuable as it establishes a disciplinary 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 additionally find value in accurately understanding and precisely characterizing the effects of and phenomena around technology. However, our community has limited resources to meet these goals, and we have historically chosen to devote more resources towards novelty and innovation while trading-off smaller sized empirical studies that are less accurate and precise, without giving up gaining understanding. A large part of our community has adopted particular statistical tools - such as frequentist null hypothesis testing - and quantitative standards, such as *p* < .05, that define this trade-off and set what constitutes sufficient evidence for researchers to support their claims and publish novel results.

This trade-off and the fast pace of CHI has lead to concerns about the precision and accuracy of knowledge accrued in our field. It is well known in the statistics community that results from individual studies -- especially with the small sample sizes typical in our community -- silently fail to reliably estimate true effects [cite ionnadis?]. To gain more reliable and precise estimates of effects, multiple studies could be aggregated using frequentist meta-analytic techniques. Such techniques combine the results from multiple papers to better evaluate claims of interest to the community. However, our community rarely practices this technique, as meta-analyses typically combine prior work without contributing new techniques or designs.

Based on the high-level ideal of accruing knowledge across studies, several calls for improving knowledge accrual have emerged. Paralleling so-called *replication crises* in psychology and medicine, movements such as RepliCHI [] have called for an increased focus on replication and meta-analysis to effectively accrue quantitative knowledge about the utility and generalizability of the results found in single studies comparing different system designs. 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" []; or at CHI, Kaptein & Robertson []), or the abandonment of frequentist null hypothesis significance testing (NHST)[[1]](#footnote-1) statistics altogether for Bayesian analyses.

All of these suggestions, apart from a Bayesian approach, are logical extensions of NHST, the dominant statistical approach used in CHI and related fields such as psychology. NHST is a statistical approach that asks if we can reject or fail to reject a *null hypothesis* 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 found to be below the customary target of .05, then the new system is deemed to be an improvement, as this *p*-value suggests that the observed difference is not due to chance: we can reject the null hypothesis of no difference.

However, getting a p-value less than .05 can still happen even when there is no true difference. Replication and meta-analysis allow us to reduce this error in NHST (e.g., 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 to be conducted and published (the meta-analysis), necessitating new top-down incentives for conducting and publishing meta-analysis in CHI. By contrast, Bayesian analysis allows us to incorporate the prior knowledge from other studies of the same and similar phenomena into a paper’s quantitative analysis. A series of papers analyzing novel contributions thus inherently accrues knowledge and bypasses the need for publishing separate meta-analyses. This allows us to increase the precision and accuracy 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 without changing the incentives or the tools amounts to blaming the users instead of the tools. Instead, we propose changing the tools --- from frequentist 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 user-centered statistics.

In the rest of this paper, we first outline why Bayesian statistical approaches are more appropriate for the CHI community than NHST. We demonstrate that Bayesian analysis better matches the CHI community’s needs because it aligns better with the actual questions the CHI community asks, gives answers that are more useful (e.g., not just *if* something is better but the magnitude and precision of that difference based on current best knowledge), and provides a mechanism for knowledge accrual that does not require a meta-analysis (thus fitting better with the incentive structure at CHI).

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 range of effect sizes from our survey of existing meta-analyses. 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 the ratio of meta-analyses to primary research is very low.
2. Bayesian analysis **provides more precise estimates of previously-studied conditions in each successive study**. The frequentist approach only increases the precision of effect sizes in a new study if that study has a larger sample size 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 to be conducted (which is anyway unlikely to be done at CHI).
3. Bayesian analysis **allows more precise comparison of novel conditions against known conditions**. By giving more precise estimates of effects of previously-studied conditions, Bayesian analysis also 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 top-down incentives for the publication of meta analyses, addressing the low ratio of meta-analyses to primary research by 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.

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

## Replication and meta-analysis in CHI

The statistical tools researchers in CHI have do not currently help them effectively accrue knowledge from one study to the next, even when the variations in design of novel systems are often informed by previous work. The classic strategy for knowledge accrual of a series of NHST studies, even in CHI, is brief literature reviews conducted often in the related work section of a CHI paper, that implicitly uses the *vote-counting* method of knowledge accumulation. 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 sample sizes to estimate statistical significance via *p*-values less than .05; in particular, many of these significant differences are likely due to chance [].

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 []. While encouraging more standalone replication studies and meta-analyses is useful for supporting knowledge accrual, we argue that it has difficulty fitting into CHI culture and the incentives for publishing novel findings. As we will show later, there is little meta-analysis currently being conducted in the community, supporting our intuition.

## Bayesian statistics versus frequentist statistics

Interpretations of frequentist statistics are a common source of errors amongst practicing researchers. The focus on p-values and significance testing amounts to insisting that users should learn how to interpret the conceptual double-negative that is a *p* value, rather than interpreting results as evidence for a hypothesis --- an interpretation only valid within a Bayesian framework. Even switching to confidence intervals from *p* values does not solve this problem, as a 95% confidence interval cannot be interpreted as an interval containing the parameter 95% of the time []. Thinking in *p*-values and frequentist confidence intervals puts an unnecessary cognitive burden on users.

By contrast, Bayesian analysis gives us a formal approach 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 a hypothesis. Prior beliefs can be derived from previous work, allowing us to accrue knowledge from study to study without requiring a separate meta-analysis to be published. 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 will we 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 --- i.e. quantitative knowledge increases in precision with each new study, rather than only accumulating when a meta-analysis is conducted. 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).

# A Survey of Effect Sizes and Meta-Analyses in HCI

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 for 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 [cite,cite,cite]. 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 systems and HICSS to specialized venues (ICMI '06: Proceedings of the 8th international conference on Multimodal interfaces). Meta-analyses are not being rewarded by the current publishing incentives of the community.

# Contrasting Bayesian and Frequentist Statistics in Estimating Effects 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: Progress bars 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.* []; thus, we can derive 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.* [] found that using a *slow-to-fast* progress indicator (i.e., one whose progress advances slowly initially and speeds up near the end of the survey) decreased the probability that a person would complete the survey. They found an effect size (as a log odds ratio[[2]](#footnote-2)) of ~-0.45.[[3]](#footnote-3) 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. 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.* []. We will also surmise a similarly-sized effect of *fast-to-slow* progress indicators, in the opposite direction, of 0.45.[[4]](#footnote-4)

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

* **Experiments 1-3** all compare 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).

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 of each world, we conduct a logistic regression in each experiment to model the probability of *completion* based on *progress indicator*: *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.* []. This yields a final, more precise estimate of the effect of that indicator based on the preceding four experiments.

### Bayesian analysis

In the frequentist 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, we build upon the previous results by using the posterior distribution 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. This has the effect of incrementally accumulating knowledge --- in the Bayesian worlds, we do not require a fifth publication, the meta-analysis.

In experiment 4, we must also place a prior on the new *slow-to-fast* indicator. We use a Cauchy[[5]](#footnote-5) distribution centered at 0 (no effect) with a scale equal to the furthest point in the 95% credibility interval 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 the control condition.

## Results

### In a single world

Before considering the differences in results across all simulated worlds, we will first walk through the results in one simulation. The results of the frequentist analysis are shown in Figure XXX.A and the results of the Bayesian analysis in Figure XXX.B. Each figure shows a 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. The dashed vertical lines indicate the true effect sizes from which the data was simulated.

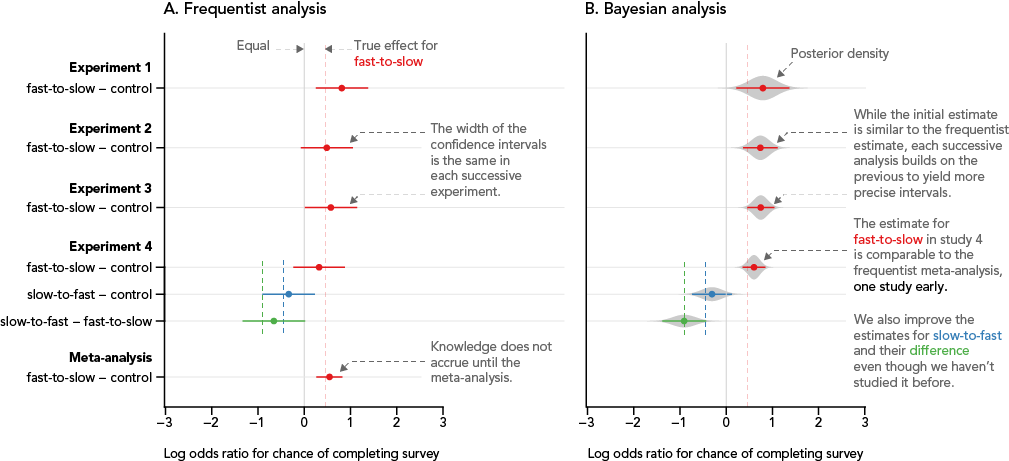


Figure XXX. Forest plots of effects from the frequentist and Bayesian analyses applied to one of our simulated worlds.

1. Learn (i.e., make a more precise estimate) faster
2. Learn (i.e., make a more precise estimate) with fewer studies
3. Can apply old knowledge to novel questions (e.g., comparing previously studied versions to new versions - insights faster)

**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.

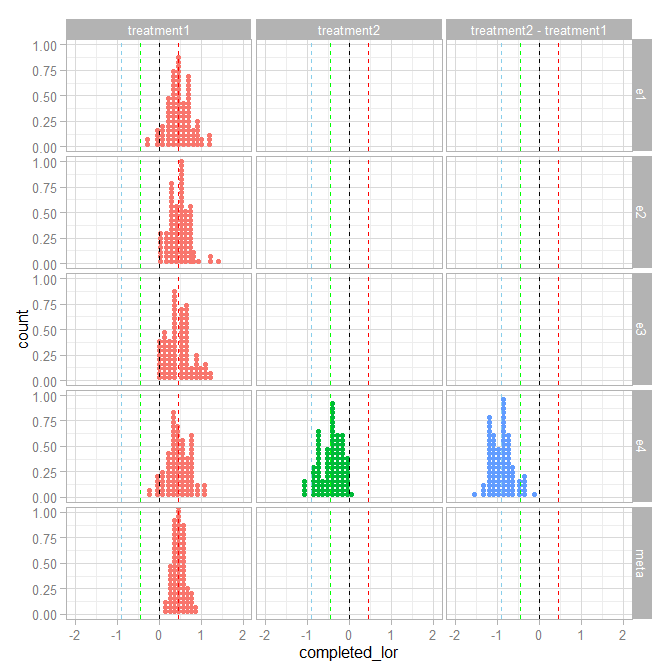
In addition, the small variation in intervals from experiments 1-3 represent vastly different conclusions if we reduce the results to null hypothesis tests: experiments 1 and 3 reject the null (p < 0.05); experiment 2 does not. This highlights the problem with reducing estimation to a binary choice (“effect” 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 happens to overlap 0.

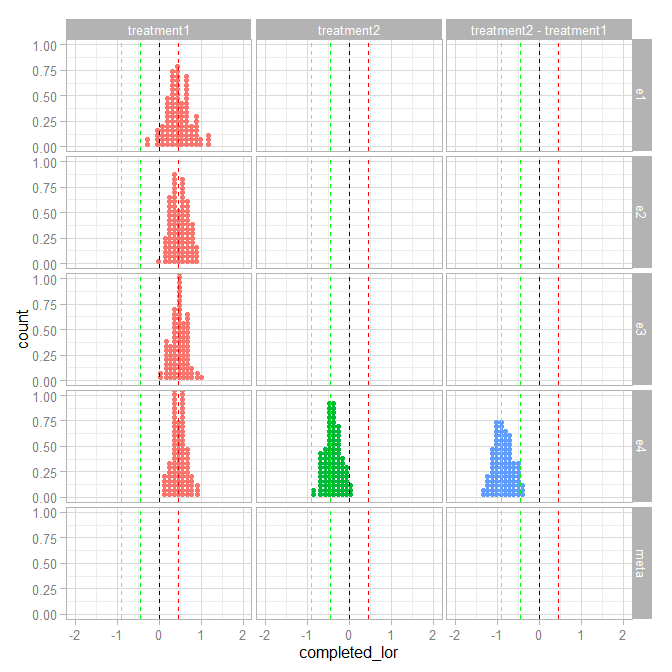
**In the Bayesian analysis,** the result of the first experiment is virtually identical to the frequentist world (we used a weakly-informed Cuachy(0, 2.5) prior for logistic regression parameters recommended by Gelman *et al.* []). However, in contrast to the frequentist world, in each subsequent experiment 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 meta-analysis, this also has additional benefits for the authors of experiment 4: note that, even though they are testing against 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 YYY.1 represents the mean estimated effect from the frequentist analysis in one of the simulated worlds. Figure YYY.2 shows the mean estimated effects from the Bayesian analyses.





[Figure YYY. As above, to be annotated etc.]

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:

> effects\_rmse(freq\_effects)

E4 RMSE for freq\_effects

treatment1 0.27

treatment2 0.27

treatment2 - treatment1 0.26

> effects\_rmse(bayes\_effects)

E4 RMSE for bayes\_effects

treatment1 0.17

treatment2 0.20

treatment2 - treatment1 0.22

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, HCI studies are often conducted with fewer participants than a traditional power analysis would suggest is prudent [cite Kaptein]. With a frequentist analysis, this increases the probability of what Gelman calls a *magnitude error* []: 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 importance of systems and engineering contributions to the field. Thus, we ask: can Bayesian analysis help make better use of our limited 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 behaviour. This knowledge is encoded in the prior we set on the effect, adapted from Gelman [].

Bayesian analysis effectively weights how strongly our prior knowledge is versus 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 encode this intuition as a prior and use it 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 Bayesian estimate:

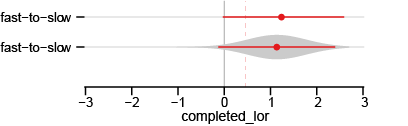


Figure XXX. Shrinkage. [needs annotation; first row is a freq estimate from one simulation of exp 1; second row is bayes estimate of same]

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’s notion of significance (note that to reach significance here the sample effect would need to be nearly 3 times the actual effect!).

### In many worlds

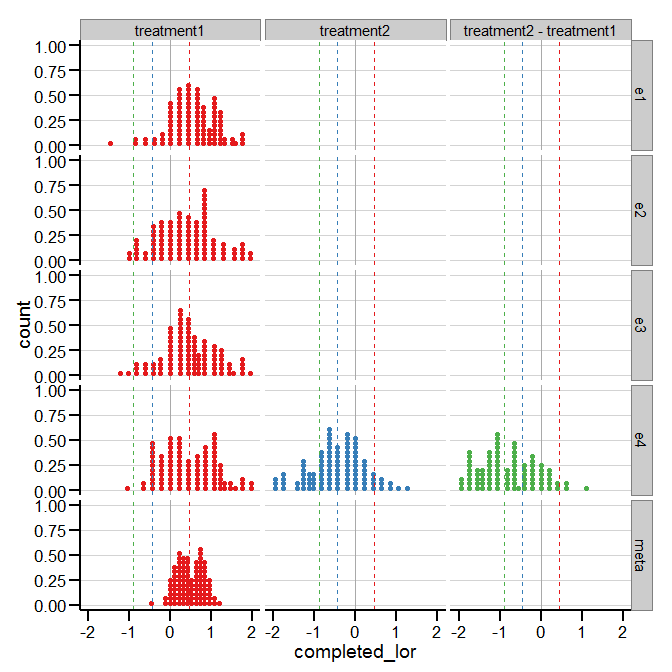
We can see the effects of shrinkage on the first experiment when we look at all of the 20-participant simulations (Figure XXX): the most extreme estimates in experiment 1 are moved slightly towards 0. This has the effect of reducing the overall error in experiment 1 by discounting unreasonably large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

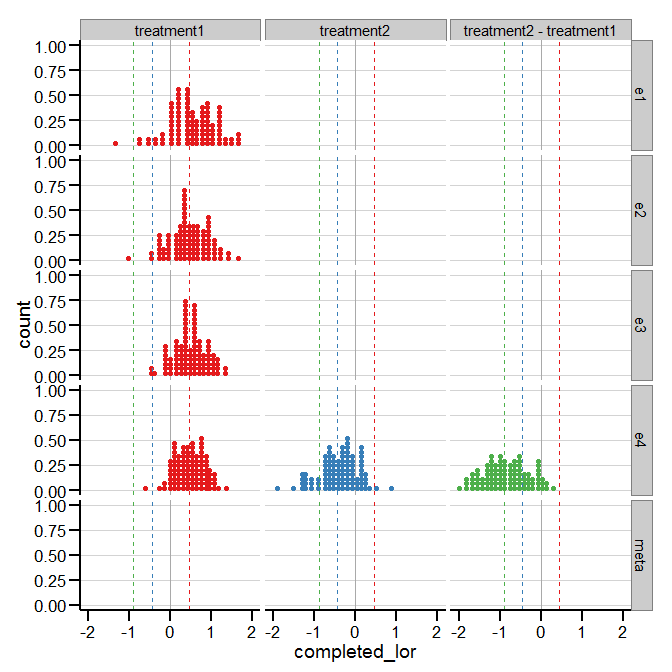
E1 RMSE for freq\_effects

e1 treatment1 0.61

E1 RMSE for bayes\_effects

e1 treatment1 0.56





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*.

E4 RMSE for freq\_effects

e4 treatment1 0.66

e4 treatment2 0.68

e4 treatment2 - treatment1 0.83

E4 RMSE for bayes\_effects

e4 treatment1 0.36

e4 treatment2 0.51

e4 treatment2 - treatment1 0.60

# 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*.

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 []); telling them not only to run evaluations but to recruit more participants amounts to blaming the users.

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 effective tool instead of blaming the user.

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” 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. Novel work should anyway (and already often is) judged on the merits of design and engineering, not a *pro forma* small-*n* evaluation. Bayesian analysis helps us make better use of these initial evaluations.

## 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.* [] 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.[[6]](#footnote-6) 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. As Leek *et al.* [] argue, 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 followup work. This is the perfect 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, a template to follow. Such a paper also provides priors for the next researcher. In this manner (we hope) such analyses will spread in the community, slowly building a body of work and a new standard of practice. We believe this to be a more likely avenue

## Bayesian analysis is accessible to practitioners

Even 15 years ago, Bayesian analysis was arguably impractical for most researchers. 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 modelling languages like JAGS [] and Stan [] (both with R packages, and Stan includes a Python interface), and Python-specific libraries like emcee [] and PyMC []. In addition, literature aimed at practicing researchers has made Bayesian modelling accessible: we particularly recommend Kruschke’s *Doing Bayesian Data Analysis* [] (which includes a table of common frequentist analyses and their Bayesian equivalents), as well as his proposed BEST test,[[7]](#footnote-7) a robust Bayesian alternative to the *t*-test []. Other accessible articles have also been written about practical concerns in Bayesian analysis, including discussions of how to choose priors [].

## 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 significance is not the language of practitioners or business; cost/benefit analysis is. The results of a Bayesian analysis are trivially 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 results into the real world, and gives us a way to put practical effect size 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 differences between conditions, will help free design and engineering researchers from the shackles of meaningless p value 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 designed for the CHI community.

# Cut from RL

* right-turn on red story
* truth inflation (statistical significance filter)
* etc

## Conf intervals and effect size estimation

* The New Statistics, Kaptein, that simulated studies paper from alt.chi
* other critiques of p values; need to find Gelman’s quote about null hypotheses being equivalent to asking where a ball is in a room and only caring about whether it’s exactly in the corner or not.
* Kruschke, Gelman
  + basically, what are all the other benefits we aren’t going to bother reiterating in the rest of the paper?
* equivalence of MLE and Bayes with flat prior as connection to what we’re doing: the old way is basically the equivalent of saying we don’t know anything given prior work; Bayes gives us a way to actually use that work directly.

Here’s what the field could look like (with meta-analysis) but it doesn’t; here’s what it could look like w bayesian. it would have the same publishing incentives. you could make these additional knowledge claims and questions. bayesian stats more understandable and is an evolution of prior work practices that meets goals of the field, subject to publishing constraints. it’s easier for users less versed in stats. we’ve linked to tools to help you adopt it.

# Other discussion points

* reporting guidelines: sourcing priors (ideally entire graph). Providing full posterior sample in supplementary material.
* Within-subejcts to reduce noise
* bridge to prior practice - (for a small domain) ask people for their data and re-analyze it with bayesian statistics and frequentist to show comparison for the field, then do hypothetical story of how would have gone if had done bayesian in the papers
* Also: how to extract priors from past work that didn’t provide bayesian results. e.g., using SE to derive variance, etc.
* statistics as lang for hci knowledge claims. here’s the things we can make claims on with freq, here’s claims we can make with bayesian and it’s wider
* frequentist as analogous to uninformative priors ; see some discussion here: <http://andrewgelman.com/2015/09/04/p-values-and-statistical-practice-2/>
* in hci everyone chooses their own way of evaluating various claims. because as a field we don’t do similar enough tasks, we are not able to get the following kinds of knowledge - H J K. We dealt with this by x y z in the paper
* maybe just make it easier by providing scripts and inputs / tasks for evaluations in a library somewhere - maybe as expected part of contributing a new area or when making review papers?
* also, even given different measures, we can use what we know about similar measures to form priors (by at least estimating expected effect sizes and using that as prior on variance). Plus, if we use this approach, we have greater incentive to use measures from previous studies, which is anyway a good thing.
* it would be nice for us to refocus contributions to not being which of these two interfaces is better but instead what elements of the interface or particular ones, and hopefully those correspond some parameters of the model, and we can use it to explain those differences and quantify how much we know about them
* address ionnadis ref about most published research is false and importance of publishing negative and other results by reframing as always contributing some information via studies when we use this bayesian framework, which shifts from “it’s different” to how different (and how precise our estimate of that difference is). mention this method helps people publish more but doesn’t address this problem fully.
* ppl wondering for interface w many features, was it this one or many, can we go back far enough to when that feature was invented, to get ideas for effects there from that evaluation, but the issue is the tasks won’t be comparable
* doing this incrementally plays in with us not reading older studies often
* getting people’s old code and rerunning studies and compare it to an estimation of what we think the results will be that we’ve built using the data available from the papers , to show how we just really can’t do that and say that means should change some things in the field
* proposing some “uniform” set of baseline tasks or problems and doing them for a bunch of prior work so that people can have priors to use within this new statistical/knowledge framework
* work that people do in studying/modeling scientific systems - greg has read some papers in this area and will do a bit more of that (we could read to just cite a bit or read to try to expand the paper, but expanding may not be necessary, wait to hear from professors when we have more, to see if expansion is needed)
* Constructing priors from meta-analysis: see discussion here: <http://andrewgelman.com/2015/08/31/constructing-an-informative-prior-using-meta-analysis/>
* Good for UISTy people
* **how/why do people use quantitative results in HCI?** Currently: they take their own quantitative results (and no one else’s) and use them as evidence for design implications *in the same paper*. The problem with this is that even if meta-analyses are conducted, they are never “between” (causally) the quantitative results and the design implications. So design implications are always based on the noisiest type of quantitative evidence that the field has. Whereas if A Bayesian approach is used where quantitative evidence in any given paper is more precise due to use of prior knowledge, then the quality of the quantitative evidence used for design implications is higher. Looked at another way, it’s a way to put evidence of the quality of meta-analyses into papers before design implications are written.
* **pilot studies**: helping to make pilot studies more robust, particularly with a focus on making far more focused predictions that can help to better define failures in design early in the process
  + **skeptical priors in small studies**: a related idea that might be worth talking about is setting "skeptical" priors so that unreasonably large effect sizes in small studies are shrunk a little closer to zero
  + **effect on large effect sizes from truly novel things (e.g. full text search, programming abstractions):** won’t be shrunk to zero, meanwhile the typical things (small-to-medium)
  + **From Eric email:** <http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124> as a great place to set that priors. You will also notice the commentary from one person suggested Bayes Theorem: <http://www.plosmedicine.org/annotation/listThread.action?root=69377>. I have personally been thinking that you could use these data to set up a higher bar for continuing work by establishing priors of success at 15% (the currently useful science based on John's work).
  + Also see <http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001747>
* **data sharing is a prerequisite for this**
* **reporting standards**
  + visual presentation example
  + citing priors
  + skeptical prior and cumulative replication probability
* value for practitioners: Bayesian analysis is in the language of practitioners: cost-benefit analysis. The language of statistical significance is not the language of business; cost-benefit analysis is.

## Objections

* we only ever study really novel things in HCI. how could we possibly have knowledge about what those things will do?
* easy response is that expected magnitudes are at least a thing. Plus, we still have to compare to previous state-of-the-art (which we should be able to get good priors for)

# Conclusion

* Here’s what the field could look like (with meta-analysis) but it doesn’t; here’s what it could look like w bayesian. it would have the same publishing incentives. you could make these additional knowledge claims and questions. bayesian stats more understandable and is an evolution of prior work practices that meets goals of the field, subject to publishing constraints. it’s easier for users less versed in stats. we’ve linked to tools to help you adopt it.

**Things to keep in mind for the paper**

* if have some room to make choices in defining the framework, steer it towards having researchers be able to publish more than they would have in the other framework (this could come in with a focus on modelling effects: even if we dont see a difference, it is valuable to say “we haven’t seen an effect but the direction so far is *x* and our precision is *y*”. Someone else can then take that and improve their future estimates. This gets at the problem of truth inflation and small-n studies in hci; if we’re gonna run small-n, underpowered studies it is actually better if we publish *more* studies with effects that aren’t credibly different.
* the paper hinges on having good visual explanations for these things. should be able to read the paper by reading the figures
  + for example, the traditional meta-analysis figure
  + visual diff of papers in both traditions
* what questions do we want to answer with HCI research
  + is this thing better
  + how much better is it
  + (probably not in this paper) what impact does that have on wider task model? ie faster scroll when spend most time reading text won’t help you much

**Overall Actions Remaining:**

* table of common HCI response types, traditional analyses, Bayesian versions (see Kruschke)
* come up with a hypothetical experiment, response variable, effect size to analyze in the simulations.
  + some areas of infovis - tasks more uniform or perceptual work
  + fitt’s law as a toy example for us to generate studies for to prototype that component before we figure out a better task / study model
* run simulations
* pilot different approaches for the above to see what people respond too
* reach out to people to pilot draft paper (jeff, jake, others?)
* (maybe) do more simulations and have them as addenda / attached materials? like for other kinds of hypotheses / claims

* (maybe) finding some data from other papers

**Other thoughts:**

* How we redo stuff in hci - including this other condition as a comparison; other thing can learn from prior studies even if no one has studied what you’re doing, what types of effects should be expected. prior can tell you how much variance should expect - prev interfaces have caused in order of X improvements, shouldn’t expect that if everything else is in that range, shouldn’t see big huge effects. Let’s you make comparisons across papers without needing access to code and running direct comparisons.
* What are the ultimate results in terms of publishability? things are more publishable? We get closer to truth faster (this could happen solely from if you have better estimates of an effect, won’t do studies where P(study publishable | true effect size) is lower, since you’d get better estimates of true effect size)? Is there some way to use it during prototyping, ie you don’t see the right size of effect when early prototyping, so you may abandon that earlier before a larger study
* modeling scientific process:
  + if someone publishes something with a large effect, will more people move to that space? or is it more dominated by prior investments in knowing prior work etc.

# Misc (orphan) points

* we already agree on a role for papers in putting together how they build on prior work. that’s why we have prior work sections. but we’re not including prior quantitative work in current practice.
* we can think of this as a user-centered design problem, where researchers are the users. what methods can we use that fit into the way people already do work in the field? the constraints are publishing incentives in the field and compatibility with prior practices so it can be adopted. the knowledge needs for the hci field are x y z (how do we establish this? reference other sciences like psychology where they say they need meta-analyses?). statistics is a language that supports a set of these needs.
* additionally, we show that this approach (positively, neutral, negatively) affects the publishability of results. (anything that helps people publish more would probably be favorable? --- worth considering weight of evidence and confidence interval arguments, plus incentives to avoid truth inflation / file drawer problem) separately, we may show benefit on the estimate of actual effect for small effects, by moving a result that from frequentist perspective isn’t significant into publishability in the bayesian framework. thus bayesian would also improve accuracy of what the field believes, since in meta-analysis, these studies wouldn’t get published and thus not included.
* As a field HCI is generally undertheorized, throw lots of work at a wall and see what sticks. We are missing opportunities as a field to build more certain knowledge. We systematically overestimate effects through publication bias. If we think a quantity is worth knowing such that we reject publications based on it’s value, is it not worth knowing accurately?

Statistical methods serve multiple purposes in HCI. They express potentially generalizable knowledge about the world.

HCI considers an increasingly wide range of phenomena, yet has adopted a comparatively narrow range of statistical tools to provide evidence for claims.

Quantitative evaluations ask questions like How much better is my system according to some metric, like task duration, completion rates, error rates, etc? Is my system different compared to another system?

* Here we can make an aside about lack of validated instruments and standard tasks. how to handle the case when the prior work’s measurements are of the same fundamental data type or scale (like from a 1-5 rating to a 1-7 rating)? We answer this criticism in the discussion.
* We’ll focus on a single scenario in this paper
  + Optional bonus: in our additional materials there are graphs / analysis for each of the kinds of analysis given in the table above

derive an estimate of effect sizes we can expect at CHI

# orphans From meta-analysis section

--Estimate of Coverage of Meta-Analysis: < XXs %

The number of papers in CHI in the dl with the quantitative keyword is X. We assume all of these used statistical methods. We then counted the number of studies each meta-analysis considered for inclusion and the number that met the inclusion criteria (this considers papers outside of CHI). Creating a precise estimate of the severity of this problem would take substantial effort, and we present this estimate to help quantify what is clearly a problem. We also note that, even if a complete meta-analysis of all work was done, not every paper could be combined together, which we discuss further in the discussion.

Figure XX reviews the effect sizes and statistical methods in a sample of these meta-analyses chosen to show a diversity of areas studied and effect sizes. In Section XXX we will use these as a basis for simulations to compare frequentist and Bayesian techniques for aggregating quantitative evidence. < we can discuss some of the interesting results here and provide a case example here of why knowing effect sizes is valuable or at least interesting. the effect size of sleep deprivation from one meta-analysis on task performance from the human factors community for example provides an interesting natural scale to interpret other results by ie this interface improves performance as much as performance is hurt from being 24 hours sleep deprived).

# Old outline of simulation section

# Incremental Meta-analyses Using Bayesian Statistics

* description of bayesian stats versus frequentist
* graphic comparing old practice not referencing prior work and new practice, showing the double reporting of just this study and this study using X other studies priors

# an illustrative comparison of bayesian and frequentist statistics in estimating effects in <domain>: a section

matt is writing this section and has simulations mostly completed.

TODO FOR THIS SECTION:

1. generate 100 examples of hypothetical worlds
2. choose 1 to write the main part of this section about.

*qualitative argument by narrative of how it would work and its benefits:*

* Write a story. imagine we did this with bayesian statistics, someone did these experiments, they can combine their effects, they learn this, they make this decision differently
* first, a simple example with same outcome, same task, different interface
* now we show how to incorporate obvious prior knowledge for a non-obvious case, with different tasks or not same baseline

<here’s a different way to maybe lay out the paper, moving more story into the intro>

* we could move this to more of the intro. first tell the story of what is currently done in some small domain we can get data for. early paper lay out questions A B C. review in text what they did and what was learned. then do a meta-analysis. show the meta-analysis is cool and what we learn about this small domain that’s not in any one paper by answering questions A B.
* now, again with a narrative approach, show how earlier papers in this small domain could have laid out questions A B C in reference to a bayesian model / framework. show what they would have done differently, show some studies could have been avoided. show that we might have noticed question C wasn’t getting answered. show how the meta-analysis results are basically the same.
* We would then move on to the simulation by saying “we showed an example, but does this example generalize for the field. for the following 1-3 models in other areas, we simulate these scientific progressions of studies. we show the bayesian framework works in these other cases and that it has the following benefits: …

<end different way>

<now present conclusions, but argued from this cherry-picked example>

* frequentist approach - even if you do best you can, giving confidence intervals instead of p values, in each study you’ll see more variability, meta-analysis fixes that but no one does the meta analysis
* then in bayesian version you’ll see priors cause subsequent studies’ effects to be more narrow; then still building but we get better estimate of the effect and don’t have to deal with the fact that there is very little incentive for publishing meta-analysis

**simulating work to compare approaches**

* That was just an example. We now simulate research progressions in hci and show these benefits are not limited to the previous example. We also show more methods and their claims.
* We present a reasonable enough model of an hci study. here’s type of experiment that is run in hci, and the assumed model of how the world looks. lets simulate 5 studies from this model, what would have happened if people had analyzed it with frequentist methods, what would a meta-analysis look like, ; the other way, imagine you feeding priors into them, argue that people should be using similar enough tasks for this to work (but later discuss that this isn’t common practice.)
* what is more likely to be publishable and how would that affect the field, ie
  + publishability
  + effect on beliefs in model parameters
  + effects on beliefs for effect sizes
  + closeness of estimates to the actual effect over time
* show simulations from a model of some task / effect in hci and show how bayesian method with priors would compare to frequentist meta-analysis at the end of k studies.
  + what dimensions will we compare these on - timeliness, ability to avoid studying poor / negative effects, effect on publishability, effect on not abandoning a research direction too early
  + what “simple” task or other model can we identify and re-use or create on our own?
  + how do we make the task concrete enough and meaningful/important enough to help bolster our argument?
  + maybe we say the first study analyzed it with one kind of outcome variable, the second had different outcome variable. with a little bit of thought, we can come up with priors that still kinda fit, … so we handle the nuance of even with people using different outcomes
* We should show how the method is robust to having evolving tasks, outcomes, and interfaces. For example, take two studies on this previous thing, but then paper 3 includes previous one and a new one, next 3 were all on that new one

1. While there exist Bayesian formulations of NHST based on Bayes factors, we consider these share some problems with frequentist NHST, such as a focus on binary testing rather than precision of estimation and cost-benefit analysis; thus we do not consider them here. [↑](#footnote-ref-1)
2. 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. [↑](#footnote-ref-2)
3. While their results use probability of drop-out, we use probability of completion. [↑](#footnote-ref-3)
4. It is worth noting that this effect may be larger than the true effect in the real world, since Villar et al. [] 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*. [↑](#footnote-ref-4)
5. 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. [] [↑](#footnote-ref-5)
6. According to Google Scholar, accessed 2015-09-23: <https://scholar.google.com/scholar?cites=16254127723353600671&as_sdt=5,48&sciodt=0,48&hl=en> [↑](#footnote-ref-6)
7. Somewhat glibly, BEST stands for *Bayesian estimation supersedes the t-test* [↑](#footnote-ref-7)