Arguing Practical Significance in Software Engineering Using Bayesian Data Analysis

Richard Torkar, Robert Feldt, and Carlo A. Furia,

Abstract—This paper provides a case for using Bayesian data analysis (BDA) to make more grounded claims regarding practical significance of software engineering research.

We show that using BDA, here combined with cumulative prospect theory (CPT), is appropriate when a researcher or practitioner wants to make clearer connections between statistical findings and practical significance in empirical software engineering research. To illustrate our point we provide an example case using previously published data. We build a multilevel Bayesian model for this data, for which we compare the out of sample predictive power. Finally, we use our model to make out of sample predictions while, ultimately, connecting this to practical significance using CPT.

Throughout the case that we present, we argue that a Bayesian approach is a natural, theoretically well-grounded, practical work-flow for data analysis in empirical software engineering. By including prior beliefs, assuming parameters are drawn from a probability distribution, assuming the true value is a random variable for uncertainty intervals, using counter-factual plots for sanity checks, conducting predictive posterior checks, and out of sample predictions, we will better understand the phenomenon being studied, while at the same time avoid the obsession with $p$-values.

Index Terms—practical significance, statistical significance, Bayesian data analysis, empirical software engineering

1 INTRODUCTION

We claim that Bayesian data analysis (BDA) can be used as a foundation to better discuss practical significance in empirical software engineering (ESE) research.

Statistics, we argue, is one of the principal tools researchers in empirical software engineering have at their disposal to build an argument that guides them towards the ultimate objective, i.e., practical significance and (subsequent) impact of their findings. Practical significance is, as we have seen [45], not very often explicitly discussed in software engineering research publications today and we argue that this is mainly out of two reasons.

The first one being that statistical maturity of ESE research is not high enough [45], leading to difficulties with connecting statistical findings to practical significance. The second reason is a combination of issues hampering our research field, e.g., small sample sizes, failure to analyze disparate types of data in a unified framework, or lack of data availability (a recent study in our field showed that only 13% of the publications provided a replication package and carefully described each step to make reproduction feasible [21]).

Both of the above issues are worrisome since it could make it hard strengthening any arguments concerning practical significance, i.e., connecting effort and, conclusively, ROI to the findings of a research study, if one would want so.

In the end, issues such as the above are likely leading ESE towards a replication crisis as we have seen in other disciplines, e.g., medicine [19], [22]–[24], psychology [1], [29], [43], economics [9], [25], and marketing [21].

In order to solve some of the above challenges researchers have proposed that we need to focus on, e.g., (i) openness, i.e., that data and manuscripts are accessible for all stakeholders, (ii) preregistration, i.e., a planned study is peer-reviewed in the usual manner and accepted by a journal before the experiment is run, so that there is no incentive to look for significance after-the-fact [12], (iii) increasing the sample size, (iv) lowering the significance threshold from $p < 0.05$ to $p < 0.005$ [4], and (v) removing null hypothesis significance testing (NHST) altogether, which the journal Basic and Applied Social Psychology advocates [46], as does the authors in [33].

Some researchers, most notably Gelman [17], claim that the above is not enough but instead a unified approach for these matters should mainly evolve around three solutions: Procedural solutions, solutions based on design and data collection, and improved statistical analysis.

Concerning procedural solutions, Gelman [17] like others, suggests publishing papers on, e.g., Arxiv, to encourage post-publication review, and to use preregistration as a tool for lowering the ‘file drawer’ bias. For design and data collection, Gelman provides convincing arguments that we should focus on reducing measurement error (the example being that reducing the measurement error by a factor of two is like multiplying the sample size by a factor of four), and move to within-subject from between-subject study designs when possible.

1. In literature, Return-On-Investment refers to, in various ways, the calculation one does to see the benefit (return) an investment (cost) has.
Finally, concerning improved statistical analysis, Gelman advocates the use of Bayesian inference and multilevel models (MLMs)3 as a way to discuss “...the range of applicability of a study”, i.e., practical significance. Overall, we side with these arguments and think they are critical also for software engineering to better connect empirical research with the practice it ultimately aims to improve.

In our paper we rely on four key concepts, which will help us establish a better understanding concerning practical significance: Bayes’ theorem, multilevel models, Markov chain Monte Carlo sampling, and cumulative prospect theory.

Bayes’ theorem states that,

\[ P(A|B) = \frac{P(B|A)P(A)}{P(B)} \]

where \( A \) and \( B \) are events, and \( P(B) \neq 0 \). In this formula we have two conditional probabilities, \( P(A|B) \) and \( P(B|A) \), the likelihood of event \( A \) occurring given that \( B \) is true, and vice versa. The marginal probability is then observing \( A \) and \( B \) independently of each other, i.e., \( P(A) \) and \( P(B) \). Often the above is rewritten as, \( P(A|B) \propto P(B|A) \times P(A) \), i.e., the posterior is proportional to the likelihood times the prior or, in other words, given a likelihood and a prior we will be able to approximate the posterior probability distribution; this is, of course, also applicable to MLMs.

Bayesian MLMs have several advantages 32: (i) When using repeated sampling they do not underfit or overfit the data to the extent single-level models do (maximally), (ii) the uncertainty across uneven sample sizes is handled automatically, (iii) they model variation explicitly (between and within clusters of data), and (iv) they preserve uncertainty and makes much data transformation unnecessary. In our particular case, this is done by using Bayes’ theorem as the foundation for conducting inference, and Markov chain Monte Carlo (MCMC) as the engine that drives it.

Hamiltonian Monte Carlo (HMC), which is one common MCMC method, uses stochastic sampling procedures to approximate integrals 5. To not spend time on tuning such a procedure by hand, we turn to Stan, which provides “a probabilistic programming language implementing statistical inference”, as the developers describe it.

Finally, we will rely on cumulative prospect theory (CPT) as a way of dealing with decision-making under uncertainty 30. Since the outcome of a Bayesian data analysis is a stochastic model of the phenomenon under study we can use it to study, i.e. simulate, different real-world scenarios. While several decision making frameworks could be used, CPT includes a more psychologically informed and up-to-date view of how people judge risks and outcomes. CPT, which partly is a development of the expected utility function 36, takes into consideration that decision makers handle risks in a non-linear fashion. CPT can be used to explain how, e.g., decision makers in industry, when faced with what empirical researchers believe to be a sure bet, instead opt for status quo, i.e., when betting on winning, people in general overweight options that are certain, and are risk averse for gains, i.e., the certainty bias. Additionally, when having to make a decision people tend to focus on the things that are different between two options to decrease the cognitive load (isolation effect), while trying to avoid losses (loss aversion effect). Mathematically speaking, the refinement of prospect theory, i.e., CPT, is formulated as 47:

\[ V(O) = \sum_{i=m}^{n} \pi_i \nu(x_i) \]

where \( V \) is the subjective value of the outcome \( O \), \( \pi \) is a decision weight (captures the idea that people tend to overact to small probabilities, and underact to large probabilities), and \( \nu \) is a function defining the subjective value of outcome \( i \).

We will next present related work in ESE where BDA and MLMs have been applied. Section 5 points to relevant literature for applying BDA in our line of research. In Section 4 we present one way to classify practical significance. This is then used on an example (Section 5), where we re-analyse previously published data to provide a delimited case making it easier to conceptually follow our line of thought. This is then explicitly connected to practical significance by using CPT (Section 6). In Section 7 we map our results from using BDA, with CPT, with the model for practical significance from Section 4. We conclude our findings in Section 8.

2 RELATED WORK

There are few publications in software engineering where we see evidence of using MLMs. In 13 the authors used multilevel models for assessing communication in global software development, while in 20 the authors applied MLMs for studying reviews in an app store. However, both studies used a frequentist approach (maximum likelihood) i.e., not a Bayesian approach.

As far as we can tell, there have only been two studies in software engineering that have applied BDA with MLMs 14, 15. In 15, Furia presents several cases of how BDA could be used in computer science and software engineering research. In particular, the aspects of including prior belief/knowledge in MLMs is emphasized. In 14, on the other hand, the author presents a conceptual replication of an existing study where he shows that MLMs support cross-project comparisons while preserving local context, mainly through the concept of partial pooling 8 as seen in Bayesian MLMs.

In our case, we will certainly use the concepts of priors and partial pooling—they are after all key to BDA and MLMs—however, our end-goal will be to connect out-of-sample predictions, i.e., one of the outputs from BDA, to the concept of practical significance, where we make predictions on new data.

Finally, concerning the combination of MLMs and cumulative prospect theory (CPT), Nilsson et al. 55 used MLMs to estimate the parameters for CPT in mathematical psychology, i.e., when an experiment has been conducted where CPT is evaluated, what are the parameters estimated from

3. Multilevel models can also be called hierarchical linear models, nested data models, mixed models, random coefficient, random-effects models, random parameter models, or split-plot designs.

4. http://mc-stan.org/

5. Partial pooling takes into account variance between units.
the experiment? As we will see we rely on such estimates for validation purposes concerning practical significance.

3 (BAYESIAN) STATISTICS AS PRINCIPLED ARGUMENT

We will not spend time on discussing the pros and cons of using statistics, but we rather assume the reader to be schooled in the benefits of using statistics, and understanding the importance statistics plays for our, and other, research fields.

Much literature on BDA exist, but not all have the clarity that is needed to explain, sometimes, relatively complex concepts. If one would like to read up on the basics of probability and Bayesian statistics we recommend [28]. For a slightly more in-depth view of Bayesian statistics we would recommend [10] (in particular Ch. 11 is well worth a read, providing arguments for and against Bayesian statistics). For a hands-on approach to BDA, we recommend [32]. McElreath’s book Statistical Rethinking: A Bayesian Course with Examples in R and Stan is an example of how seemingly complex issues can be explained beautifully, while at the same time help the reader improve their skills in BDA. Finally, there is one book that every researcher should have on their shelf, Bayesian Data Analysis by Gelman et al. [18], which is considered the leading text on Bayesian methods.

In this paper we will show that using BDA will allow us to better connect to practical significance. But we first need to articulate what practical significance is in our context.

4 PRACTICAL SIGNIFICANCE

A review has reported that less than 50% of the publications, from five top journals in software engineering, explicitly discuss practical significance [45].

We believe it would help the software engineering community if researchers would (i) be expected to explicitly report on practical significance in their studies, (ii) not be allowed to only make implicit arguments for practical significance based solely on the nature of the problem/context, e.g., “Company A thought this was important”, and (iii) required to clearly connect any findings with own or others’ evidence. However, it is also worthwhile to keep in mind that there exist software engineering research that does not immediately connect to practical significance, but where it is rather seen as a long-term goal.

We postulate that an analysis of practical significance should ideally be built around the following maturity model:

1) Context. Identify the practical contexts in which the found effect is important. The context typically includes which type of domain, company/organization type, size, experience level, etc. where the study took place or where the results apply. See [38] for a detailed model of context, and [7] for an elaboration on the importance of context via-à-vis generalizability.

2) Affected variables. Identify which outcomes would be affected by the effect in the chosen contexts. Outcomes are typically high-level predictors like cost, effort/time, risk, or quality, but can also be more concrete metrics related to the top-level ones.

3) Absolute practical significance. Argue for why the size of the effect, as shown by the statistical analysis, would have practical significance for the identified variables in the given contexts. The maturity of this argument is based on what type of evidence exists, e.g., (from lower maturity to higher)

   a) reason from common sense. Researchers assume it is evident to the reader once the effect has been stated or refers to grey literature opinions about relevant levels.
   b) compare to published literature. Compare effects to relevant variable levels as supported by empirical data in published literature.
   c) static interpretation. Assessing the evidence, this is mainly done by interpreting how important the effect of the identified size is. This should preferably be done by using statistics as principled argument.
   d) dynamic interpretation. Complementing the static interpretation by, i.e., questioning practitioners in the relevant contexts, and present evidence of their interpretations of how important the effect of the identified size seen in their context is. This should preferably be done by using statistics as principled argument.

4) Relative practical significance. Argue for how the seen effect(s) fares concerning alternative methods or a change in the affected outcomes/predictors. The argument here can be at the same o–d refinement levels used in #3 above.

We argue that absolute practical significance (Item 3) does not make sense and cannot adequately be described if one has not identified the variables in Item 2. And they, in turn, depend on the context identified in Item 1. Although an explicit discussion of the (absolute) practical significance in Item 3 can often be considered enough, there are always alternative solutions that can be selected. The analysis in Item 4 aims to clarify how sensitive the established effects are compared to alternative solutions, or when other variables (incl. hypothetical) are used. The above model can be seen as a general maturity model for arguments about practical significance in empirical software engineering since lower, earlier levels are a pre-requisite for higher, later levels.

The proposed maturity model is general and applies regardless of how practical significance is argued. However, by using BDA, and, in particular, by combining it with CPT, we believe it will be easier for researchers to discuss absolute and relative practical significance. Thus, while the maturity model puts up the goals to be achieved, BDA together with a decision making support framework like CPT is a concrete way to achieve the goals. We will next present a case of how this analysis and argument can be done.

5 A CASE FOR BAYESIAN DATA ANALYSIS

Generally speaking, the three main steps of Bayesian model design and analysis are:
1) Understand the data and the problem.
2) Design a probability model (conduct model checking and iterate if the model needs to be revised), and sample from the posterior to conduct diagnosis.
3) Conduct inference. That is, learn something about the population by using a sample of said population, e.g., by conducting statistical tests or deriving estimates.

The above is an iterative process, and in the last step we also have the possibility to change the parameters to see how they affect the outcome variable, i.e., to potentially analyze practical significance. The above steps will next be covered in our example.

5.1 The Data and the Problem

The data we use in this analysis has partly been published before [2]. The data is from an experiment to understand the effects of using exploratory testing, where 35 subjects participated. Of the 35, 23 subjects were classified as less experienced (LE) and 12 were classified as more experienced (ME). The experiment evaluated two techniques, i.e., a new technique (NT) and an old technique (OT), used a small, noncritical system as the software under test, and had a 2 × 2 design to avoid learning bias. NT is exploratory testing, while OT is traditional test case based testing. The effectiveness of each technique was measured by true positives (tp), i.e., the number of faults the technique found that were classified as true faults. Below we see the first rows of the data file that we will use (2 observations/subject adds up to 70 rows):

```r
> head(d)
subject category technique tp
1  1  LE    NT  20
2  1  LE    OT  1
3  2  LE    NT  4
4  2  LE    OT  1
5  3  LE    NT  9
6  3  LE    OT  0
```

We would like to understand if the new technique (NT) is better than the old technique (OT), and if there is a difference between less and more experienced subjects (LE/ME). This way we will be able to decide if the technique should be used by a company and, hence, if there is a need to take experience levels into consideration. The original experiment showed that there was no learning bias introduced.

The original study found that NT was significantly better than OT, and that there was no significant difference between LE and ME. Other studies have also partly confirmed these results, e.g., [26], [27]. For further details on the original experiment we refer you to [2].

5.2 Design of Model and Diagnosis

In BDA it is generally considered good practice to test our assumptions and spend time on setting proper priors for our parameters, i.e., conducting a sanity check, before fitting a data generative model to draw samples from the posterior. Hence, we need to (i) create a first basic model, (ii) draw parameter values from this model, (iii) generate data from these values, (iv) fit the model to the generated data, and (v) check the fit. In short, a model should have a good fit using the data it generates.

Let us start with a simple model with additive terms, i.e.,

\[
\begin{align*}
tp & \sim \text{Poisson}(\lambda) \\
\log(\lambda) & = \alpha + \beta_t \times \text{technique} + \beta_c \times \text{category} \\
\alpha, \beta_t, \beta_c, \beta_s & \sim \text{Normal}(0, 1)
\end{align*}
\]

In the above model we assume \( \text{tp} \) to be from a Poisson distribution with an event rate \( \lambda \), and we then write out our linear model with the logarithm (log) as the link function \( \log(\lambda) \) We also set generic weakly informative priors, i.e., we simply say that we do not expect very extreme values like infinity, for the intercept \( \alpha \), and for each of the parameters \( \beta_t, \beta_c, \beta_s \). Finally, we fit the model using no outcome values, only sample from the priors, and check that the chains mix well.

Next, we investigate if \( R \approx 1.0 \) (the potential scale reduction factor on split chains, which should approach 1.0 at convergence), that the effective sample size is acceptable (should not be too low), and that the parameter values are not too wide. In the latter case, we notice that \( \alpha \) (the intercept) has a 95% uncertainty interval of \([-36; 36]\), which is too extreme (the other parameters are approximately CI\%50% \([-2; 2]\)). This is a strong indication that the priors are too wide—remember that we are using only the priors now and that we disregard the likelihood.

To analyze the output of too wide parameter values we need to, possibly, (i) provide more conservative priors, i.e., since four additive terms, with each having a prior of \( \mathcal{N}(0, 1) \), leads to the prior for log intensity to be \( \mathcal{N}(0, 4^2) \). In addition, we should introduce (ii) multilevel features to capture the variability of the \text{subject} predictor, and (iii) introduce the concept of zero-inflated distributions to emphasize that there are 18.6% zeros in our outcome variable, \( \text{tp} \). (In the end, we will see that Stan handles our generic weakly informative priors well.)

The sanity checks we performed indicated the importance of checking assumptions and conducting an analysis of priors, which ultimately led us to the following model:

\[
\begin{align*}
tp_i & \sim \text{ZIPoisson}(p_i, \lambda_i) \\
\logit(p_i) & = \alpha_p + \beta_p \times \text{technique} \\
\log(\lambda_i) & = \alpha + \beta_t \times \text{technique} + \beta_c \times \text{category} + \alpha_{\text{SUBJECT}[i]} \\
\alpha_{\text{SUBJECT}} & \sim \text{Normal}(\mu_s, \sigma_s) \\
\alpha & \sim \text{Normal}(0, 10) \\
c(\alpha_p, \beta_p, \beta_t, \beta_c, \mu_s) & \sim \text{Normal}(0, 1) \\
\sigma_s & \sim \text{HalfCauchy}(0, 1)
\end{align*}
\]

First, we state that the outcome, \( \text{tp} \), comes from a generative process that follows a zero-inflated Poisson distribution. On the next line we model the probability of zeros, \( p_i \), as a linear model depending on the \text{technique} predictor. Then, we have another linear model, which contains \( \alpha \) the intercept, and the two parameters \( \beta_t \) and \( \beta_c \) for \text{technique} and \text{category}, respectively. \( \alpha_{\text{SUBJECT}[i]} \) is the

6. A link function provides a relationship between the linear predictor and the mean of the distribution function.
Let us next look at the effects of introducing ‘multi’ in our multilevel model, i.e., we model our subjects as having varying intercepts, and we assign two hyperparameters, $\mu_s$ and $\sigma_s$, to be part of the prior for $\alpha_{\text{SUBJECT}}$, see, e.g., [32] for an explanation. Finally, we assign priors for the rest of our parameters, while for $\sigma_s$ we settle on a weakly regularizing prior for standard deviations, i.e., HalfCauchy$(0, 1)$, which only allows positive values.

The diagnostics show that the chains mix well, in addition, the effective sample size is high, and $R \approx 1.0$ indicates the chains have converged. Figure 1 confirms that our original outcome ($y$) seems to be fairly inline with 100 random draws ($y_{\text{rep}}$) from the posterior probability distribution, which represent our simulated outcomes. Figure 2 provides another view where we have plotted our fitted values, $y_{\text{rep}}$, against our actual responses, $y$. There seems to be a linear fit, albeit $y$ contains higher values; this is an effect of introducing partial pooling to decrease the risk of overfitting, in combination with using a zero-inflated distribution.

Let us next look at the effects of introducing ‘multi’ to our model, i.e., modeling the subject predictor using varying intercepts and hyperparameters. By introducing multiple levels in a Bayesian model we, in principal, try to avoid under- and overfitting, that is, learning too little or too much from our data.

Generally speaking we here discuss three main ways to deal with this type of uncertainty (i) no-pooling, i.e., each subject’s ability to find a true positive is completely independent; none of the information from different subjects are pooled together (a separate regression for each subject), (ii) complete pooling, i.e., all information is shared; we fit one line and disregard that there were 35 subjects involved (grand mean), and (iii) partial pooling, i.e., the subjects come from a population and we estimate the properties of that population through each subject’s ability to find a true positive. In short, we estimate the priors with hyperparameters that are themselves estimated.

If we compare two plots of central (quantile-based) posterior interval estimates from our draws—partial pooling and complete pooling—it is clear that partial pooling adapts to the information at hand (Figure 3). For partial pooling, the estimates we receive for our subjects provide less underfit than in the complete pooling estimates, while still being able to provide less overfit than the no-pooling estimates, i.e., where we treat each subject as unique and do not learn anything from the information we have retrieved from the other subjects. However, we do pay a price for this, i.e., our intervals are larger, but embracing that uncertainty is something we are willing to do.

5.3 Conduct Inference

The inferences will provide us with posterior probability distributions that captures uncertainty and beliefs concerning the process we are trying to understand better, i.e., how well a technique works in finding defects taking into account different levels of experience among staff.

If we examine Table 1 (a partial visual presentation is seen in Figure 4) we see that both technique ($\beta_1$) and category ($\beta_2$) seem to be significant, i.e., zero is outside the 95% uncertainty region of both parameters; however, category is borderline. Recall that the original study did not find a statistically significant difference for the category predictor, which seems to be the opposite case now.

Let us next examine a pairs plot where we have univariate marginal distributions along the diagonal, as kernel density plots, and bivariate distributions off the diagonal.
The first row provides the standard deviation for the group effect (subject). The next three rows of estimates is the intercept (α), and the main effects of technique (βt) and category (βc), respectively. The last two rows of estimates are from the zero-inflated part. The 95% uncertainty intervals (UI) indicate that βt and βc are not crossing zero, and the std. error for both parameters are, in the worst case (βc), not even half of the effect’s estimate.

Fig. 4. Density estimates of βc (top) and βt (bottom). The 95% probability density of βc comes close to crossing zero.

(Figure 5). As we can see there is a negative correlation between the α, i.e., the intercept, and βc, i.e., the parameter for the predictor category, but that only tells us that the two parameters hold some of the same information. When building more complex models these types of correlations can be problematic and we could see chains that do not mix well and divergent transitions, i.e., the sampler could get stuck in a local minima. In our particular case we have good mixtures, sane R values, and no divergent transitions, so we will leave it for now.

We should also take the time and investigate marginal effects in our model. A marginal effect is the expected instantaneous change in the outcome as a function of a change in a certain predictor, while keeping all covariates constant. Figure 6 shows clearly that something happens when we change the levels in our predictors, in particular concerning the predictor technique.

Finally, we have one outstanding issue to take into account, i.e., the ‘significant’ finding that level means actually do make a difference (see βc in, e.g., Figure 5 and Table 1). As we pointed out earlier, the original article [2] showed that there was no effect, while our analysis indicates that there might be one after all. In order for us to deal with this uncertainty it would be wise to further analyze this effect. To this end we apply ROPE analysis, i.e., region of practical equivalence [31], as a way for us to estimate which effects are practically relevant for us to consider.

By using the posterior probability distribution, together with the uncertainty connected with said distribution, we investigate the relation to a ROPE around the null value.

In the social sciences one often use Cohen’s δ, i.e., the difference between the means of two things being compared \((\bar{x}_1 - \bar{x}_2)\), divided by the standard deviation of the sample \(s\). A small effect is often signified as \(\delta = 0.2\), and thus an equivalent rule for a ROPE analysis, as Kruschke writes [31], would mean ROPE \(\pm 0.1\), or for a very small effect ROPE \(\pm 0.005\) (in [42] the author then argues for \(\delta = 0.01\)).

As we see in Table 2 a ROPE analysis of βc shows that we cannot reject the hypothesis that βc is outside the ROPE interval. In this particular case we opted for ROPE = 0.1, thus we tested for a small effect. Hence, we conclude that even though βc has a credible interval not crossing zero, the region of practical equivalence indicates that the effect is, perhaps, not that significant. However, that viewpoint might not necessarily hold when we connect this to practical significance, i.e., a risk/effect, albeit small, can have a large implication.

### 6 Connecting to Practical Significance

To connect our BDA to practical significance we need a measurement that is relevant to us. We will provide a decision maker with > 1 choices regarding the introduction of a new technique and the different factors guiding said introduction. In short, we need to calculate the utility of introducing a possible solution by using CPT and connect it to our BDA where we have posterior distributions for our estimates.

| Estimate | Est. Error | UL95% |
|----------|------------|-------|
| 0.28     | 0.09       | 0.10  |
| 1.96     | 0.10       | 1.75  |
| -1.43    | 0.17       | -1.77 |
| 0.31     | 0.15       | 0.01  |
| -3.72    | 0.95       | -5.85 |
| 2.77     | 1.15       | 0.65  |

| TABLE 1 |

Fig. 5. Pairs plot of the parameters of interest. We can see that there are indications of negative correlations between the parameters, in particular between α and βt.

| H0 | % in ROPE | HDI95% |
|----|-----------|--------|
| α  | reject    | 0.00   | 1.77  |
| βt | reject    | 0.00   | -1.77 |
| βc | undecided | 0.64   | 0.01  |
| zi α | reject | 0.00   | -5.67 |
| zi βt | reject | 0.03   | 0.59  |

| TABLE 2 |

By using the posterior probability distribution, together with the uncertainty connected with said distribution, we investigate the relation to a ROPE around the null value.

In the social sciences one often use Cohen’s δ, i.e., the difference between the means of two things being compared \((\bar{x}_1 - \bar{x}_2)\), divided by the standard deviation of the sample \(s\). A small effect is often signified as \(\delta = 0.2\), and thus an equivalent rule for a ROPE analysis, as Kruschke writes [31], would mean ROPE \(\pm 0.1\), or for a very small effect ROPE \(\pm 0.005\) (in [42] the author then argues for \(\delta = 0.01\)).

As we see in Table 2 a ROPE analysis of βc shows that we cannot reject the hypothesis that βc is outside the ROPE interval. In this particular case we opted for ROPE = 0.1, thus we tested for a small effect. Hence, we conclude that even though βc has a credible interval not crossing zero, the region of practical equivalence indicates that the effect is, perhaps, not that significant. However, that viewpoint might not necessarily hold when we connect this to practical significance, i.e., a risk/effect, albeit small, can have a large implication.

### 6 Connecting to Practical Significance

To connect our BDA to practical significance we need a measurement that is relevant to us. We will provide a decision maker with > 1 choices regarding the introduction of a new technique and the different factors guiding said introduction. In short, we need to calculate the utility of introducing a possible solution by using CPT and connect it to our BDA where we have posterior distributions for our estimates.
We can now offer decision makers two scenarios from our statistical analysis, i.e., introduce the new technique or not, and if one chooses to introduce the new technique, decide if one should generally speaking aim to increase the number of more experienced subjects in the teams (at a cost?) Visualizing the choices (Figure 7) is easy since they stem directly from our BDA (see Appendix A for details).

Each scenario, τ (introducing a new technique) and ε (increasing the experience level of the staff), has two decision points, A and B. A decision maker needs to decide which decision point is the most favorable to her. If we take ε as an example, we see that, for A, there are three probabilities, \( p = 0.03 \), \( p = 0.94 \), and \( p = 0.03 \), and a cost or loss associated with each probability (here in $).

The cost/loss for each probability are calculated using artificial data in this case, since these numbers can be sensitive for a company to release (our discussions over the years, with various companies, indicate clearly that these numbers vary a lot). In the two examples we approximate hourly salary to be $100 and $200 for less and more experienced staff (average of $134.30 for the current staff composition), we also take into account that a session is four hours, and we set the cost of a fault slipping to release to be $150 (this was a small, noncritical system). We would like to emphasize that these numbers would, of course, vary a lot depending on domain, company, etc., as has also been pointed out in, e.g., [6]. The point, however, is that a researcher can estimate such values based on reason, prior literature, or by directly talking to practitioners in the studied context. In fact, a more detailed analysis could introduce uncertainty also in these probabilities and cost/loss estimates so that the simulation of the scenarios could provide uncertainty intervals for outcomes.

For each scenario’s decision points (A and B) we can now calculate the prospect theory utility: \( \tau_A = 415 \), \( \tau_B \approx 21 \), \( \epsilon_A \approx 219 \), and \( \epsilon_B = 233 \). Hence, we expect decision makers to prefer the new technique over the old (\( \tau_A \gg \tau_B \)), but we also see that the scenario where one should choose between introducing more experienced subjects in the teams, does not give the same convincing answer. Here we should expect decision makers to not prefer the cost (and efficiency) of more experienced subjects (\( \epsilon_B > \epsilon_A \)), something also indicated by our ROPE analysis. But what happens if the cost of a released subject is significantly higher, say, $500? In Figure 8, we see this scenario.

In the new scenario the stakes are higher, and what is surprising is that decision makers would switch their preferences, since \( \epsilon_A = 1311 \) and \( \epsilon_B = 1006 \), i.e., \( \epsilon_A \) is now the preferred choice.

Before moving to the next section, let us summarize our analysis of practical significance. First, BDA provided us with a posterior distribution that captures the uncertainty given our actual observations/data. We concluded by noting the effect category had on the outcome. Second, our ROPE analysis gave at hand that the effect was small and probably not important for us to take into account in

7. For details on the calculations we refer you to the Appendix A.
Further analysis. Finally, using a decision support framework such as CPT, and simulating different scenarios and their corresponding costs, we saw that by varying, e.g., the cost of a fault slipping through, it allowed us to gain a better understanding of the threshold values important for us. Overall, we can see that the more detailed statistical analysis and the uncertainty kept in the final answers can be utilized in further discussions of and arguments for the practical significance. This is in contrast to the results of the original study being more binary, leaving a reader to remember simply that ‘exploratory testing was better’. The refined level of detail is much more informative and allows a more detailed representation of knowledge.

7 Discussion

We have argued that Bayesian data analysis combined with simulation of different, practical scenarios and decisions allows a more detailed and refined understanding of actual effects as well as their implications. Since empirical software engineering ultimately aims to have impact on the actual work of engineers, giving concrete advice for what an analysis of practical significance should provide and how it can be achieved, is important. While our proposed, simple, maturity model for practical significance provides the former, Bayesian data analysis combined with a decision making framework such as CPT can provide the latter.

The type of ‘what-if’ scenarios we have presented are very valuable during discussions with engineers in industry, and compared to the current modus operandi in our community: “βc and βt were statistically significant with p < 0.05 using an arbitrary α = 0.05”, a leap forward in interpreting what significance means from a practical point of view. However, developing these scenarios required us to use BDA, which contributes with a posterior probability distribution that captures uncertainty as well as co-variation between the considered factors and effects.

In Section 4, we postulated that an analysis of practical significance should entail at least four steps. The first two steps, Context and Affected variables we presented in Section 5.1 and Sections 5.1 and 5.2, respectively.

Regarding the third step, i.e., Absolute practical significance, we argue that a dynamic interpretation, i.e., δ(d), has been achieved; however, not through discussions with practitioners in industry, but rather by relying on CPT, which has been validated repeatedly and ultimately led to Kahneman receiving the Nobel Memorial Price in Economics in 2002.

The usage of BDA in combination with CPT also allowed us to conduct a dynamic interpretation of the Relative practical significance, by showing how changes in variables affected the practical outcome of our analysis.

Nevertheless, what we have presented here is merely scratching the surface of how one can use BDA and Markov chain Monte Carlo for more realistic inferences. Since we have a posterior probability distribution we have the possibility to change predictors as needed, or introduce new predictors altogether, while constantly capturing the ever-changing uncertainty; for example, conducting simulations, such as, “How will our outcome be affected by hiring new employees”, is quite straightforward in BDA (see Figure 9).

We will not further contrast this approach with how analyses are done in empirical software engineering today. Suffice to say, issues such as the arbitrary α = .05 cut-off, the usage of null hypothesis significance testing (NHST) and the reliance on confidence intervals has been criticized [23], [34], [37], [49], and when analyzing the arguments, the authors of this paper have concluded that many of the issues plaguing other scientific fields are equally relevant to come to terms with in empirical software engineering.

8 Conclusions

By making a case for BDA we showed that, through the use of CPT, a richer analysis can be reached allowing us to gain a better understanding of the phenomenon we study. We argue that, by following this approach, understandability and usefulness will increase, mainly by simulating various scenarios and embracing uncertainty.

The BDA we conducted showed partly different results compared to the original study [2]. First, there seemed to be a difference between less and more experienced subjects in the experiment. Second, a ROPE analysis indicated that the difference was not large. Third, CPT, in combination with the output from our Bayesian analysis, provided us with a deeper understanding of when this effect mattered from a practical point of view.

Our hope is that the reproducibility package accompanying this paper will allow researchers to (i) analyze how we founded our arguments in the analysis, and (ii) criticize and find flaws in our approach.

We expect that future work will be focused on more complex analyses, where BDA will show even more benefits. In addition, we expect that significant effort needs to be spent on theoretically aligning CPT with BDA, to better understand practical significance in connection to BDA.
Appendix A

Reproducibility Package

Begin by installing Docker and ensure that you have enough RAM and CPU assigned. Then execute the following in a terminal (remove the backticks around `pwd` if this is executed on Windows),

```
docker run -d -p 8787:8787 -v "pwd":/home/rstudio/working -e PASSWORD=YOUR_PASSWORD -e ROOT=TRUE torkar/docker-b3
```

Finally, start your browser, enter http://localhost:8787 and then use `rstudio` as login and the password you set above. In the browser you now have RStudio running, make sure to load the script `brms.R`, which you will find in the `files` window. All plots will be displayed in RStudio when executing the script.

The script and data file can also be downloaded at: https://github.com/torkar/docker-b3; however, to ensure reproducibility we recommend users to use the Docker image.

Acknowledgments

We express our thanks to Paul-Christian Bürkner, Aki Vehtari, and Jonah Gabry, for the discussion on priors, and for pointing us to literature discussing the choice of priors, i.e., [16], [44].

References

[1] Aarts, A. A., et al., 2015. Estimating the reproducibility of psychological science. Science 349 (6261).
[2] Afzal, W., Ghazi, A. N., Itkonen, J., Torkar, R., Andrews, A., Bhatti, K., Jun 2015. An experiment on the effectiveness and efficiency of exploratory testing. Empirical Software Engineering 20 (3), 844–878.
[3] Au, G., 2014. pt: An R package for Prospect Theory. The University of Melbourne, Victoria, Australia.
[4] Benjamin, D. J., et al., 2018. Redefine statistical significance. Nature Human Behaviour 2, 6–10.
[5] Betancourt, M., Jan. 2017. A conceptual introduction to Hamiltonian Monte Carlo. ArXiv e-prints.
[6] Boehm, B., Basili, V. R., Jan. 2001. Software defect reduction top 10 list. Computer 34 (1), 135–137.
[7] Briand, L. C., Bianculli, D., Nejati, S., Pastore, F., Sabetzadeh, M., 2016. The role of the tester’s knowledge in exploratory software testing. Empirical Software Engineering 20 (3), 844–878.
[8] Bürkner, P.-C., 2017. brms: An R package for Bayesian multilevel models using Stan. Journal of Statistical Software 80 (1), 1–28.
[9] Camerer, C. F., et al., 2016. Evaluating replicability of laboratory experiments in economics. Science 351 (6280), 1433–1436.
[10] Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A., 2017. Stan: A probabilistic programming language. Journal of Statistical Software 76 (1), 1–32.
[11] Cohen, J., 1988. Statistical Power Analysis for the Behavioral Sciences. Lawrence Erlbaum Associates.
[12] Dutilh, G., Vanderkervrode, J., Ly, A., Matzke, D., Pedroni, A., Frey, R., Rieskamp, J., Wagenmakers, E.-J., Apr 2017. A test of the diffusion model explanation for the worst performance rule using preregistration and blinding. Attention, Perception, & Psychophysics 79 (3), 713–725.
[13] Ehrlich, K., Cataldo, M., 2012. All-for-one and one-for-all? A multi-level analysis of communication patterns and individual performance in geographically distributed software development. In: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work. CSCW ‘12. ACM, New York, NY, USA, pp. 945–954.
[14] Ernst, N. A., 2018. Bayesian hierarchical modelling for tailoring metric thresholds. In: Proceedings of the 15th International Conference on Mining Software Repositories. MSR ’18. IEEE Press, Piscataway, NJ, USA, p. pp.
[15] Furia, C. A., Aug. 2016. Bayesian statistics in software engineering: Practical guide and case studies. ArXiv e-prints.
[16] Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., Gelman, A., Sep. 2017. Visualization in Bayesian workflow. ArXiv e-prints.
[17] Gelman, A., 2018. The failure of null hypothesis significance testing when studying incremental changes, and what to do about it. Personality and Social Psychology Bulletin 44 (1), 16–23.
[18] Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., Rubin, D., 2013. Bayesian data analysis, 3rd Edition. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.
[19] Glick, J. L., 1992. Scientific data audit—A key management tool. Accountability in Research 2 (3), 153–168.
[20] Hassan, S., Tannahamthavorn, C., Bezem, C.-P., Hassan, A. E., Sep 2017. Studying the dialogue between users and developers of free apps in the Google Play Store. Empirical Software Engineering.
[21] Hunter, J. E., 2001. The desperate need for replications. Journal of Consumer Research 28 (1), 149–158.
[22] Ioannidis, J. P. A., 2005a. Contradicted and initially stronger effects in highly cited clinical research. JAMA 294 (2), 218–228.
[23] Ioannidis, J. P. A., 08 2005b. Why most published research findings are false. PLOS Medicine 2 (8).
[24] Ioannidis, J. P. A., 06 2016. Why most clinical research is not useful. PLOS Medicine 13 (6), 1–10.
[25] Ioannidis, J. P. A., Stanley, T. D., Doucouliagos, H., 2017. The power of bias in economics research. The Economic Journal 127 (605), F236–F265.
[26] Itkonen, J., Mäntylä, M. V., Apr 2014. Are test cases needed? replicated comparison between exploratory and test-case-based software testing. Empirical Software Engineering 19 (2), 303–342.
[27] Itkonen, J., Mäntylä, M. V., Lassenius, C., May 2013. The role of the tester’s knowledge in exploratory software testing. IEEE Transactions on Software Engineering 39 (5), 707–724.
[28] Jaynes, E. T., 2003. Probability theory: The logic of science. Cambridge University Press.
[29] John, L. K., Loewenstein, G., Prelec, D., 2012. Measuring the prevalence of questionable research practices with incentives for truth telling. Psychological Science 23 (5), 524–532.
[30] Kahneman, D., Tversky, A., 1979. Prospect theory: An analysis of decision under risk. Econometrica 47 (2), 263–291.
[31] Kruschke, J. K., 2018. Rejecting or accepting parameter values in Bayesian estimation. Advances in Methods and Practices in Psychological Science.
[32] McElreath, R., 2015. Statistical rethinking: A Bayesian course with examples in R and Stan. CRC Press.
[33] McShane, B. B., Gal, D., Robert, C., Tackett, J. L., Sep. 2017. Abandon statistical significance. ArXiv e-prints.
[34] Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., Wagenmakers, E.-J., 2016. The fallacy of placing confidence in confidence intervals. Psychonomic Bulletin & Review 23 (1), 103–123.
[35] Nilsson, H., Rieskamp, J., Wagenmakers, E.-J., 2011. Hierarchical bayesian parameter estimation for cumulative prospect theory. Journal of Mathematical Psychology 55 (1), 84–93.
[36] von Neumann, J., 1947. Theory of games and economic behavior. Princeton University Press.
[37] Petersen, K., Wohlin, C., 2009. Context in industrial software engineering. Springer, Berlin, Heidelberg.
[38] R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
[39] Gelman, A., 2018. The failure of null hypothesis significance testing when studying incremental changes, and what to do about it. Personality and Social Psychology Bulletin 44 (1), 16–23.
[40] Robert, C., 2007. The Bayesian choice: From decision-theoretic foundations to computational implementation. Springer Texts in Statistical Science. Taylor & Francis.
[41] Rodríguez-Pérez, G., Robles, G., González-Barahona, J. M., 2018. Reproducibility and credibility in empirical software engineering: A case study based on a systematic literature review of the use of statistical methods in software engineering. In: Proceedings of the 15th International Conference on Mining Software Repositories. MSR ’18. IEEE Press, Piscataway, NJ, USA, p. pp.

8. https://www.docker.com
of the SZZ algorithm. Information and Software Technology 99, 164–176.

[42] Sawilowsky, S. S., 2009. New effect size rules of thumb. Journal of Modern Applied Statistical Methods 8, 597–599.

[43] Shanks, D. R., et al., 04 2013. Priming intelligent behavior: An elusive phenomenon. PLOS ONE 8 (4), 1–10.

[44] Simpson, D. P., Rue, H., Martins, T. G., Riebler, A., Sørbye, S. H., Mar. 2014. Penalising model component complexity: A principled, practical approach to constructing priors. ArXiv e-prints.

[45] Torkar, R., Feldt, R., de Oliveira Neto, F. G., Gren, L., 2017. Statistical and practical significance of empirical software engineering research: A maturity model. CoRR abs/1706.00933.

[46] Trafimow, D., Marks, M., 2015. Editorial. Basic and Applied Social Psychology 37 (1), 1–2.

[47] Tversky, A., Kahneman, D., Oct 1992. Advances in prospect theory: Cumulative representation of uncertainty. Journal of Risk and Uncertainty 5 (4), 297–323.

[48] Vehtari, A., Gelman, A., Gabry, J., 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Statistics and Computing 27, 1413–1432.

[49] Woolston, C., Feb. 2015. Psychology journal bans P values. Nature 519 (7541), 9.