Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 26 Number 4, February 2021 ISSN 1531-7714

Beyond p-values: Using Bayesian Data Analysis in Science Education Research

Marcus Kubsch[1], *IPN - Leibniz-Institute for Science and Mathematics Education*
Insa Stamer, *IPN - Leibniz-Institut for Science and Mathematics Education*
Mara Steiner, *Leibniz-Institut for Science and Mathematics Education*
Knut Neumann, *IPN - Leibniz-Institut for Science and Mathematics Education*
Ilka Parchmann, *IPN - Leibniz-Institut for Science and Mathematics Education*

In light of the replication crisis in psychology, null-hypothesis significance testing (NHST) and *p*-values have been heavily criticized and various alternatives have been proposed, ranging from slight modifications of the current paradigm to banning *p*-values from journals. Since the physics education research community often relies on quantitative statistical approaches, the challenges the replication crisis poses to these approaches need to be considered. *p*-values suffer primarily from the fact that they carry little information by themselves and lend themselves to misinterpretations. As one alternative, Bayesian approaches have become increasingly popular as the posterior distributions they provide carry more relevant information than *p*-values. In this paper, we discuss practical issues related to *p-values* with respect to interpreting and communicating results and how these issues can be addressed using a Bayesian approach. Drawing on a science education data set, we demonstrate how Bayesian data analysis methods go beyond p-values and can help to make more valid conclusions and to communicate them more easily in a manner that lends itself to less misinterpretations.

Introduction

Since its development in the early 20th century, the concept of *p*-values in frequentist null hypothesis significance testing (NHST) has been criticized by numerous authors based on theoretical and practical issues (Cohen, 1994; Gigerenzer et al., 2004; Ioannidis, 2005; McShane et al., 2017; Meehl, 1967; Simmons et al., 2011). However, *p*-values are still pre-dominant in psychological and more specifically science education research today (e.g., all papers involving statistical analysis published in *Physical Review Physics Education Research* in 2016 used *p*-values). Recently, concerns with NHST and *p*-values have resurfaced as part of the discourse about the reproducibility crisis in psychological and social science (Nuzzo, 2014; Open Science Collaboration, 2015) with well documented high profile replication failures as in Carney et al. (2010). Some researchers even go so far as to argue that low replication rates are to be expected given the current statistical paradigm (Smaldino & McElreath, 2016). Various alternatives to current practices have been proposed in the past. Some within the frequentist framework suggest modifying the current practice with *p*-values, e.g., justifying the necessary *p*-value levels in order to claim statistical significance for each study

[1]We thank Elisabeth von Bressensdorf for language editing.

depending on design, sample size, etc. (Lakens et al., 2018) or lowering the conventional significance level from 0.05 to 0.001 in order to reduce the rate of false-positives (Benjamin et al., 2018). Others go beyond current practices and argue for a "new statistic" that shifts the emphasis to parameter estimation (Cumming, 2014). Some argue for an even more radical shift and propose Bayesian approaches (Gigerenzer et al., 2004; Kruschke, 2013; Kruschke & Liddell, 2017; Wagenmakers et al., 2018).

Edwards et al. (1963), Cohen (1994), and Gigerenzer et al. (2004) have provided discussions about theoretical, mostly epistemic, differences between frequentist and Bayesian approaches. More practical discussions have been provided in the context of structural equation models (SEM) and developmental research (van de Schoot et al., 2014) or hypothesis testing and general psychology (Kruschke, 2013; Wetzels & Wagenmakers, 2012). In this paper, we demonstrate how Bayesian approaches can be useful in the context of science education research. Drawing on linear models as one of the dominant statistical tools in the field, we show how a Bayesian approach allows the incorporation of strong theoretical knowledge or knowledge from previous studies into statistical analyses and thus can help with interpreting their data. Further, given how people (including researchers) struggle with correctly interpreting the *p-values* predominant in the frequentist paradigm (Aczel et al., 2017; Gelman, 2013; Gelman & Stern, 2006; Gigerenzer et al., 2004; Wasserstein & Lazar, 2016), we discuss to what extent a Bayesian approach can support researchers in correctly interpreting statistical results as well as communicating these results more transparently.

In order to do so, we walk through the data analysis of a study in which we investigated how students' and scientists' perceptions of typical practices of scientists differ. We will provide theoretical background as necessary and present a broader discussion at the end.

An Applied Example

Background

The sciences face relatively high drop-out rates at the university level (Brinkworth et al., 2009). Among others, this may be caused by students having unrealistic ideas about what scientists do (Sharkawy, 2012; Solomon et al., 1994). In order to investigate to what extent students have realistic ideas about what scientists do, we investigated which activities high school students, graduate students, and science professors considered typical for scientists to engage in on a regular basis.

In order to do so, we administered a questionnaire to high school students, graduate students, and science professors which asked them to what extent they considered that scientists engage in activities such as "reading a journal article" on a regular basis. The questionnaire, which was iteratively developed through a number of pilot studies in order to ensure validity and reliability, used a Likert-scale ranging from one to indicate that an activity is considered untypical to four to indicate that an activity is considered typical (see Figure 1 for an example question). The activities cover a range of dimensions (**R**ealistic, **I**nvestigative, **A**rtistic, **S**ocial, **E**nterprising, **C**onventional + **N**etworking) (Dierks et al., 2014; Wentorf et al., 2015) that represent the whole span of activities that scientists engage in, e.g., social (S) covers activities such as advising a graduate student whereas investigative (I) covers activities such as conducting a literature review. For more details on the questionnaire see (Stamer et al., 2019).

Sample

A total number of 347 persons participated in the study (244 high school students, 92 graduate students, and 10 professors). The high school students aged 16 on average (*M*=16.4, *SD*=2.1) came from nine urban and sub-urban schools in northern Germany. The graduate students and professors came from a range of different physical science departments, e.g., computational chemistry or astrophysics, from a number of different German universities. All scales showed sufficient reliability for the three groups (average Cronbach's α = .76).

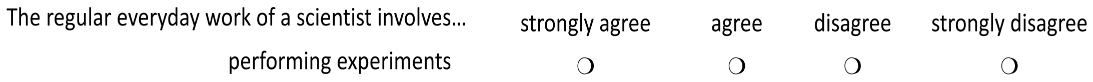
Analysis

In order to investigate to what extent high school students, graduate students, and professors have different ideas about what scientists do, we looked at the three groups on the different RIASEC+N dimensions. First, we calculated students' average score across the questions of the respective dimension. The number of questions per dimension are 4 (S), 6 (R). and 7 (C). Table 1 shows the mean and standard deviation of these scores for the three groups on the C, S, and R dimension. Now, the question arises with what confidence we can consider the differences and similarities between the groups as real, i.e., we want to know to what extent the measured differences and similarities between the groups are due to random variation or can be expected to generalize to other high school students, graduate students, and professors. How sure are we that high school students and graduate students really have different ideas about the activities on the R dimension? Are high school students and graduate students similar enough on the C dimension to be considered equivalent? How confident can we be in the small difference between graduate students and professors on the R dimension given the small number of professors? In order to answer these and similar question, we apply statistical procedures.

The current standard: Using *p*-values

Let us formulate a statistical model for our data. In this paper, we will rely on linear models for analysis as they allow us to use a consistent manner to describe the statistical models that we will apply and make differences between the approaches easily visible (Cumming, 2014; Kruschke & Liddell, 2017). We first consider the question about the extent to which high school students and graduate students are really different on the R dimension. We denote the data for the R dimension R_i. The subscript i indicates the ith participants score. As in a t-test, we assume that the scores on the R dimension for high school and graduate students taken together are approximately[2] normally distributed $N(\mu_i, \sigma)$. Figure 2 shows the close alignment of the data for R and a superimposed normal distribution which justifies the assumption of normality. In $N(\mu_i, \sigma)$, μ_i denotes the mean score on R and σ the respective standard deviation

Figure 1. Example question from the realistic (R) dimension.



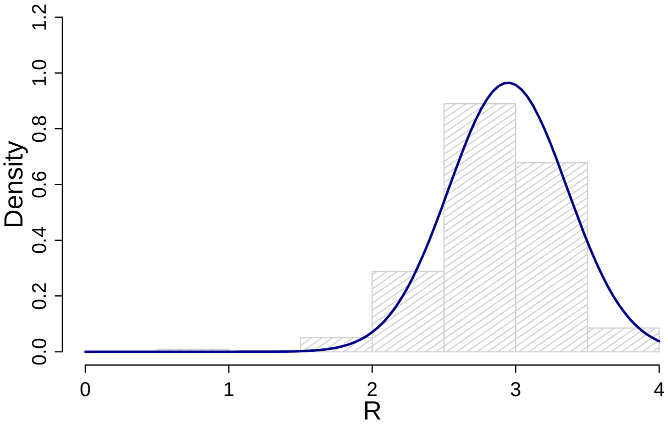
The regular everyday work of a scientist involves...	strongly agree	agree	disagree	strongly disagree
performing experiments	❍	❍	❍	❍

Table 1. Mean and standard deviation for the groups on the C, S, and R dimension

Dimension	Group	Mean	SD
C	High School Students	3.30	0.44
	Graduate Students	3.30	0.49
	Professors	3.27	0.72
S	High School Students	2.52	0.57
	Graduate Students	2.62	0.63
	Professors	3.65	0.43
R	High School Students	2.98	0.44
	Graduate Students	3.32	0.49
	Professors	2.90	0.67

[2] Please note that we make some approximations in this analysis for the sake of simplicity and refer to e.g., Gelman et al. (2012) or McElreath et al. (2016) for a thorough coverage of Bayesian Data Analysis and Likelihood functions beyond the normal distribution.

Figure 2. Histogram plot of data for R with superimposed normal distribution in black



In order to get at the difference between the groups, we write it as a linear relationship for μ_i where β_0 represents the mean of the graduate students, β_1 the difference in means between the two groups, and the dummy variable $High\ school\ student_i$ indicates whether a participant is a high school or graduate student: $\mu_i = \beta_0 + \beta_1 \times High\ school\ student_i$. All taken together, this gives the statistical model in equation (1):

$$R_i \sim N(\mu_i, \sigma) \qquad (1)$$
$$\mu_i = \beta_0 + \beta_1 \times High\ school\ student_i$$

Note that the model assumes that the groups have similar standard deviations. If this assumption is violated, we have to extend the model in order to account for different standard deviations. We use the statistical software R (R Development Core Team, 2008) in order to estimate the model, which provides us with t-values, degrees of freedom, and *p*-values. Conventionally, the question whether there is a difference between the groups will be answered based on the *p*-values[3]. In case of this model, the *p*-value for β_1 is of interest as β_1 describes the difference in means between high school and graduate students. Running the model in the statistical software R gives *t*(334)=-6.15, *p*<.001. But what does that *p*-value tell us? It gives the probability for the difference in means taking the calculated *t*-value or a more extreme one given the null-hypothesis that the true parameter difference in means is zero. In other words, the *p*-value tells us how (in)compatible the data are with our statistical model and the respective null-hypothesis of the difference between the groups being zero (Cohen, 1994; Gigerenzer et al., 2004; Kruschke & Liddell, 2017; Wasserstein & Lazar, 2016). In this case, the *p*-value tells us that our data are very incompatible with the null-hypothesis of the difference between the groups being zero. As the value is below the conventional *p*=0.05, we conclude that the difference between the groups is statistically significant and interpret our results as evidence that there really is a difference between high school and graduate students regarding the R dimension of RIASEC+N. However, statistical significance does not imply practical importance because small effects of no practical importance can be highly statistically significant given large samples (McShane & Gal, 2017). In order to judge the practical importance of the difference, we draw on effect size measures such as Cohen's *d* which in our case takes the value of *d*=0.75, 95% CI [0.51, 1.00], indicating a

[3] Note that when we want to compare multiple groups, *p*-values need to be corrected in order to account for multiple testing. In the Bayesian framework, we can address this issue using multi-level modelling techniques (see Gelman et al., 2012).

medium to large difference (Cohen, 1992). In sum, in this case, *p*-values and the effect size indicate a practically meaningful difference between the groups that we would consider real.

Let us turn to high school and graduate students one the C dimension. Both groups have similar means and standard deviations, but are they similar enough to be really considered equal? Running a linear model gives *t*(334)=-0.4, *p*=.69, *d*=-0.05, 95% CI [-0.29, 0.19]. From *p > 0.05,* we infer that there is no evidence to reject the null-hypothesis that the difference between the groups is zero and conclude that there is no statistically significant difference between the groups. Further, *d* tells us that the observed difference between the groups is of little practical importance. However, we cannot consider this evidence that there is no difference between the groups because our statistical power could just have been too small to detect a difference (Cohen, 1994; Gigerenzer et al., 2004; Kaplan, 2014; Kruschke, 2013; Daniël Lakens, 2017). When we consider high school and graduate students on the S dimension, we face a similar situation with *t*(334)=-1.35, *p*=.18, *d*=-0.16, 95% CI [-0.40, 0.08]. Again, the *p*-value provides no evidence for a difference between the groups and the observed difference of *d*=-0.16, 95% CI [-0.40, 0.08] would still be considered of little practical importance. However, if we compare the results from the S and C dimension, we see hugely different *p*-values and effect sizes but are left with the conclusion of no evidence for difference: In comparison, the data for C appear to indicate that there is no difference between the groups more strongly than the data for S which might even suggest a small – although practically minor – difference between the groups. However, *p*-values are not very helpful in helping us navigate these cases.

The small amount of information a *p*-value provides becomes even more pronounced when we consider the difference between graduate students and professors on the R dimension: *t*(10.10)=-1.92, *p*=.08, *d*=-0.64, 95% CI [-1.3, 0.02][4]. The *p*-value provides no evidence for a difference between the groups but *d*=-0.64, 95% CI [-1.3, 0.02] suggests that we observed a medium to large effect favoring the graduate students. The large observed difference is not statistically significant because two factors limit the precision on the estimation: 1) our sample is relatively small as we only have data from *N*=10 professors and 2) we had to modify our statistical model in order to account for the difference in standard deviations in the two groups. We did this by adding another linear term for the standard deviation σ which leads to the model in equation (2).

$$\begin{aligned} R_i &\sim N(\mu_i, \sigma_i) \qquad (2) \\ \mu_i &= \beta_0 + \beta_1 \times \mathit{graduate\ student}_i \\ \sigma_i &= \beta_2 + \beta_3 \times \mathit{graduate\ student}_i \end{aligned}$$

If we report this result, we can argue that we observe a big difference and that the *p*-value approaches the conventional level of statistical significance. We can even point to data from our pilot studies that suggests that professors on average score close to three which is close to the average we observed in the present study (mean = 2.90) and in general answered very similarly. Considering the evidence holistically as the ASA's statement suggests (Wasserstein & Lazar, 2016) indicates that there really is a difference between graduate students and professors regarding the R dimension. However, the literature also points out that we are prone to overestimating the magnitude of effects and potentially even get the sign of the effect wrong if we face low statistical power, i.e., in situations with small samples as in ours (Gelman & Carlin, 2014). In sum, the information *p*-values provide about the data remain somewhat inconclusive.

In this section, we presented one case where *p*-values provided evidence for a difference between groups. Bases on the *p*-values and the effect size, we were able to conclude that there is a medium to large difference between high school and graduate students with regard to the R dimension. Further, we presented two cases with data from the S and C dimension where *p*-values provided no evidence for difference between

[4] Note that the decimals in the degrees of freedom come from the Welch's t-test that was used to obtain the *p*-value. This modified *t*-test accounts for the differences in variance between the groups and as indicated in the statistical model in equation 2.

groups and the observed effect sizes suggested small practical differences. In these cases, we could not conclude that there really is no difference between the groups as our sample might have been too small to detect an effect. Alarmingly, Aczel et al. (2017) found that taking $p > 0.5$ as evidence for no difference between groups is actually happening quite frequently in psychological journals. A subsequent re-analysis revealed that in many of those cases the available evidence for no difference between the groups was quite weak which might very well be the case with our results for the S dimension where $p = 0.18$ and $d = -0.16$, 95% CI [-0.40, 0.08] could hint at a small effect which we were not able to detect with the present sample. Within the frequentist framework, we could draw on equivalence testing to address this problem as equivalence testing provides evidence whether the difference between the groups is in a specified range around zero in which differences between the groups can be considered practically irrelevant. However, this would lead us to a new methodology and set of software packages which is beyond the scope of this paper. In the last case we presented, the *p*-value provided no evidence for difference, the effect size was in the medium to large range, and we had information from a pilot study that supported the medium to large effect size. However, given the small sample, we also faced the risk of overestimating the magnitude and sign of the effect, which again resulted in a situation where *p*-values were not providing helpful information. Within the frequentist framework, we could try to incorporate the information from the pilot study by testing a directed hypothesis, i.e., that graduate students on average report higher values on the R dimension than professors. However, this would not use the full data available from the pilot study and directed hypothesis tests are rarely accepted in science education research. An alternative would be to draw on advanced methods and try to incorporate the information from the pilot study in our statistical model through penalization. Similar to equivalence testing, the latter would demand new methodology and software packages.

Besides the problems we faced when making inferences because *p*-values did not provide us with much information, we also navigated a number of potential pitfalls when interpreting the *p*-values. In the first case where we compared high school and graduate students on the R dimension, we found $p < .001$. While the respective effect is conventionally considered to be highly significant, the magnitude of the effect is not, because *p*-values are not a measure of effect size (Wasserstein & Lazar, 2016) although they are commonly misinterpreted as such (Gelman, 2013; Gigerenzer et al., 2004; McShane et al., 2017; McShane & Gal, 2017; Wagenmakers et al., 2018). Further, we were careful not to interpret *p*-values as probabilities for a hypothesis being true. When we compared graduate and high school students on the S and C dimension, we were only able to conclude very little from the relatively high *p*-values. A common misinterpretation is to interpret *p*-values as the probability of the null-hypothesis being true (McShane & Gal, 2017). This misinterpretation may very well be the cause for the instances where $p > 0.05$ was considered evidence for no effect that Aczel et al. (2017) found in psychological journals.

Before we will explore a Bayesian approach in the next section, let us consider again the last case where we compared professors and graduate students on the R dimension where our *p*-value bases analysis provided only inconclusive information, which was due to our small sample and the lack of a way to incorporate prior available information into the statistical model. Science educations' research often faces the same challenge. There is a body of substantive theory but researchers struggle to incorporate it into statistical analyses and often face small samples that do not allow for precise parameter estimation (Kaplan, 2014; McNeish, 2016; van de Schoot et al., 2014).

Going Beyond *p*-values: Bayesian Data Analysis

In this section, we will introduce the general idea of Bayesian data analysis and repeat the analyses we did in the last section in a Bayesian framework. The Bayesian perspective interprets probability as the information about uncertainty, i.e., probability quantifies the (un)certainty of our information about some aspect of the world (De Finetti, 1992). Bayes' theorem forms the basis for updating one's prior information about something based on data, i.e., considering prior information one learns from the data. Let us see how this works when we reconsider high school and graduate students on the R dimension of RIASEC+N again. Equation (3) shows Bayes' theorem.

$$\frac{Prior \times Likelihood}{Average\ Likelihood} = Posterior \qquad (3)$$

The prior is the probability distribution that describes our information about the world before we have seen the data, i.e., in our case, information about any differences between high school and graduate students on the R dimension we know from the literature or earlier studies. The likelihood summarizes the information about the world we find in the data, i.e., in our case, we use a normal distribution in order to summarize the participants' scores on the R dimension, just as we did in the statistical model in Equation 1. It is often the most influential part in a Bayesian model. The posterior is the mathematical consequence of a given prior and likelihood as specified in Bayes' theorem. It is the probability distribution that describes our information about the world given the data that we have observed, i.e., in our case, it is the information about differences between high school and graduate students on the R dimension that results from the combination of prior information and data.

Let us specify our statistical model for the difference between high school and graduate students on the R dimension in Bayesian terms. Since we summarize the data the same way, the likelihood remains the same as in equation (1).

Now we need to specify a prior for each parameter in the model (σ, β_0, β_1), i.e., our prior information about differences between high school and graduate students on the R dimension. For the specification of a prior, we can often draw on the literature or earlier studies. However, in this case, we have no prior information available. In such cases, we can specify what is often called (Gelman et al., 2014; McElreath, 2016b) a weakly informative prior, i.e., a prior distribution that considers basically all results equally likely but is skeptical of very extreme results and encodes natural constraints of the model. Let us first consider a prior for the standard deviation $\boldsymbol{\sigma}$. Popular prior distributions for standard deviations include uniform distribution and half-cauchy (McElreath, 2016b). For simplicity[5], let us consider a uniform distribution which considers all values in the specified range equally likely. We know that by definition σ is positive and, based on our previous experiences with the instrument, we consider $Uniform(0,1)$ (all values between zero and one are equally likely) a sensible prior for the standard deviation. Now, we need a prior for β_0. In our statistical model, β_0 represents the average score on the R dimension of the graduate students. Based on the range of the scale (1-5) and following the usual assumption in linear models that coefficients are normally distributed, we use a normal distribution centered at 2.5 with a standard deviation of 0.75: $N(2.5,0.75)$. The standard deviation of 0.75 represents that we do not have a lot of information about how graduate students score on the R dimension reflects the range of the scale. Finally, we need a prior for β_1 that describes our prior information about the difference between the groups. Again, we use a normal distribution but as we have no prior information about the difference between the groups, we center it at zero and use a standard deviation of 0.5: $N(0, 0.5)$. This describes that we consider a large range of differences between the groups equally likely. At the same time, the model is slightly skeptical of very large differences as those are not compatible with the range of the scale. The complete model is shown in equation (4):

$$\begin{gathered} R_i \sim N(\mu_i, \sigma) \qquad (4) \\ \mu_i \ = \ \beta_0 + \ \beta_1 \times High\ school\ student_i \\ \sigma \sim Uniform(0,1) \\ \beta_0 \sim N(2.5,0.75) \\ \\ \beta_1 \sim N(0,0.5) \end{gathered}$$

Notice how the model is identical to the one presented in equation (1) except for the priors. If we had chosen priors that consider all parameter values

[5] Please note that the specification of priors is a very important part of Bayesian data analysis that requires much care. The priors in this paper present rather simple choices that illustrate the principles. We recommend https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations for advice on prior choices.

equally likely, i.e., priors that include no information, our model would be equivalent to the one in equation (1). Thus, from a practical perspective, we could think of frequentist models as a special case of Bayesian models where we specify no prior information whatsoever.

With our statistical model at hand, we can now estimate the model. We use the open source probabilistic programming language *Stan* (Carpenter et al., 2017) which we access through an R interface called rethinking (McElreath, 2016a). Stan estimates these models using Markov-Chain Hamiltonian Monte-Carlo sampling. While it is beyond the scope of this paper to explain the sampling method and Markov-chains in detail (see e.g., McElreath (2016b) for an accessible explanation), there are two ways to assess whether the sampling was successful: visual inspection of the Markov-chains and the Gelman-Rubin convergence criterion $\widehat{R}$ (Gelman et al., 2014). Visual inspection of the Markov chains is a simple but powerful way to assess to what extent the Markov-chains have mixed and reached stationary distributions. Usually, trace plots of the chains should look like "Hairy Caterpillars" (Figure 3) when the chains have mixed and reached a stationary distribution. $\widehat{R}$ shoul be one to indicate convergence. If $\widehat{R}$ is above one, the Markov-chains usually have not converged. Values of $\widehat{R}$ above 1.01 warrant caution (McElreath, 2016b). For all models that we ran for this paper, visual inspection of the Markov-chains and $\widehat{R}$ indicated that the chains mixed and reached stationary distributions.

When we estimated the model following the frequentist paradigm, our software provided us with t-values, degrees of freedom, and *p*-values. In the Bayesian approach, the software provides us with posterior distributions for every parameter. How do we use these in order to address the question about the extent to which high school students and graduate students are really different on the R dimension? We calculate the posterior distribution of Cohen's *d* which is displayed in Figure 4 below.

Figure 3. Traceplot of a "Hairy Caterpillar" Markov-chain depicting 3000 samples. The gray box represents the first 1000 samples, which are for warm-up

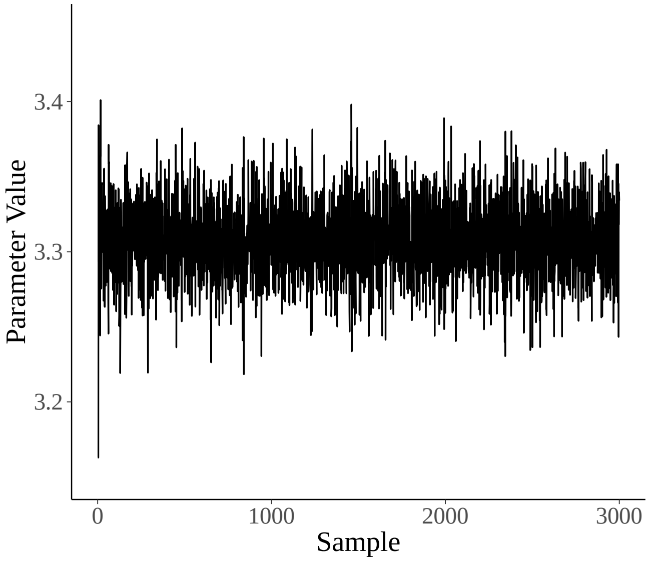
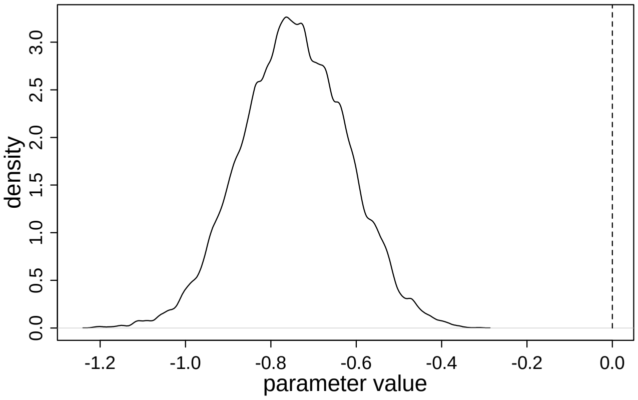


Figure 4. Posterior density plot of Cohen's d for the difference between high school and graduate students on the R dimension. Gray area marks 95% probability interval



The posterior describes the probability density of Cohen's *d* for the difference between graduate and high school students. Using the conventional 95% threshold, the posterior distribution tells us that there is a 95% probability that the difference between the groups ranges between a Cohen's *d* of -1 and -0.5 with the highest probability (the mean of the distribution) being -0.74. Thus, we can conclude that there really is a difference between the groups within the range of a medium to large effect favoring the graduate students. In this case, the results of a Bayesian approach and a *p*-value based approach are very similar. In cases where we have little to no prior information and sufficient sample sizes, this is generally to be expected. However, the plot of posterior of *d* directly communicates the magnitude of the effect and the confidence which we can have in the results. The sharper the peak of the posterior, the more confident we can be of a result and the wider the distribution the less confident we can be of a result. While *p*-value based confidence intervals (CIs) in principle can carry similar information, they are more prone to misinterpretation because CIs do not necessarily carry distributional information, i.e., while parameter values closer to the peak of the posterior distribution are more likely than those at the fringes, parameter values in the middle of a CI are not necessarily more likely than those at the fringes of the CI.

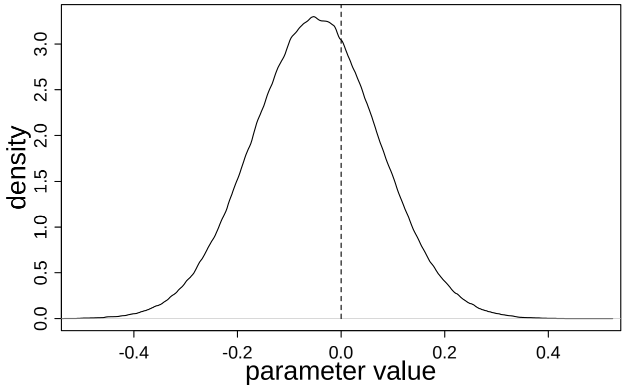
Let us now turn to the comparisons of graduate students and high school students on the C and S dimension where we faced *p*-values > 0.05 and thus were left with somewhat inconclusive results. In both cases, we have little prior information to guide us in the specifications of our priors. Thus, we will use priors similar to those used in Equation 4 and only modify the prior for β_0 which describes the average score of the graduate students to be centered around the observed mean of the graduate students on the respective dimensions[6]. Figure 5 shows the posterior of *d* for the C dimension. The posterior distribution tells us that there is a 95% probability that the difference between the groups ranges between a Cohen's *d* of -0.27 and 0.18 with the highest probability being -0.05. Further, we see that the posterior is relatively symmetrically distributed around zero. Our *p*-value based analysis of the data did not allow us to conclude that there was no difference between the groups and applying equivalence testing would have led us to a new methodology and new software tools. In a Bayesian approach however, it is quite easy to apply the idea of equivalence testing because the posterior already provides all the information we need.

[6] C: $\beta_0 = N(2.5, 0.75)$, S: $\beta_0 = N(2.5, 0.75)$

We can define a range of practical equivalence (ROPE)[7] (Kruschke, 2013; Daniël Lakens, 2017), i.e., a range within which differences between groups have no practical consequences, although they may be measurable, e.g., in an educational setting, anything below the average teacher effect of $d = 0.3$ (Hattie, 2009) may be considered a small or negligible effect. Following this definition, we can define a range of practical equivalence ranging from a Cohen's d of -0.3 to 0.3 and calculate with what probability the true difference between the groups lies in that range. We find that around 98% of probability mass of the posterior fall into the ROPE. Thus, we can infer that there is a 98% probability of practical equivalence between high school and graduate students regarding the C dimension of RIASEC+N. Thus, the Bayesian approach easily allows quantifying support for the null-hypothesis of two groups being identical using a range of practical equivalence[8].

Let us consider the difference between graduate and high school students on the S dimension. Figure 6 shows the posterior of d for the S dimension. The posterior distribution tells us that there is a 95% probability that the difference between the groups ranges between a Cohen's d of -0.30 and 0.08 with the highest probability being -.16. Equivalence testing shows around 87% of probability mass of the posterior fall into the ROPE ranging from $-0.3 < d < 0.3$. In our p-value based analysis of the S and C dimension p-values provided little information in general and also were not very helpful in distinguishing between the two cases. The Bayesian analysis however makes it easy to get a better sense of the extent to which the results provide evidence that high school and graduate students do not differ on the C and S dimension. On the C dimension, the Bayesian analysis provides evidence for practical equivalence between the groups, in case of the S dimension, the evidence for equivalence is substantially weaker. When it comes to communicating our results, the two distinctly different posterior distributions in Figure 4 and Figure 5 clearly communicate that the evidence for difference between the groups is more pronounced in case of C dimension compared to the S dimension. Because p-values are often (mis)interpreted in a p >.05 means "no evidence"[9] and p <.05 means "evidence" fashion (Gelman, 2013; McShane & Gal, 2017), there is a substantial risk that the results for the C and S dimension would both have simply been interpreted as "no evidence for difference". Thereby, the existing difference in the confidence of this interpretation for the C and S dimension would have been missed.

Figure 5. Posterior density plot of Cohen's d for the difference between high school and graduate students on the C dimension. Gray area marks 95% probability interval



[7] Ideally, a ROPE should be defined before any analysis is conducted in order to avoid bias. Also note that ROPEs will vary between disciplines, area of research etc. What matters is a sound argument for the range of the ROPE.

[8] While beyond the scope of this paper, so called Bayes factors can also be used to quantify support for null hypotheses in Bayesian data analysis (Aczel et al., 2017; Wagenmakers et al., 2018; Wetzels & Wagenmakers, 2012).

[9] While some disciplines use different thresholds for p-values, the principle remains the same.

Figure 6. Posterior density plot of Cohen's d for the difference between high school and graduate students on the S dimension. Gray area marks 95% probability interval

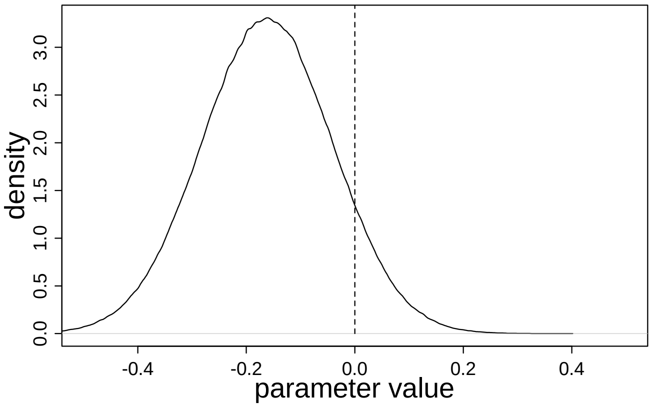
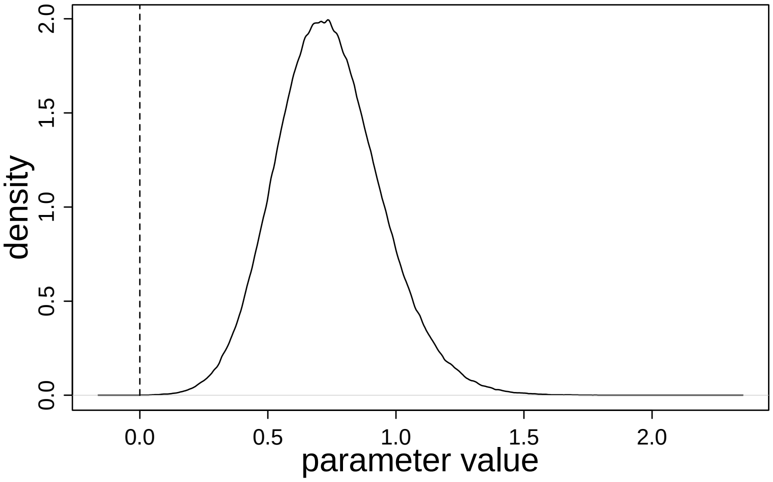


Figure 7. Posterior density plot of Cohen's d for the difference between graduate students and professors on the R dimension. Gray area marks 95% probability interval



Let us now return to the comparison of graduate students and professors on the R dimension. Using *p*-values, we faced the problem that the *p*-value was only approaching the conventional level of statistical significance while Cohen's *d* suggested a medium to strong effect. Thusly, the results were relatively inconclusive. We attributed this to the small sample and the fact that we also had to estimate standard deviations for both groups. Available prior information did not help addressing the issue because we had no way to incorporate the information into the statistical model. Now, the Bayesian approach allows us to incorporate this information into the statistical model. In a pilot study, we already saw that professors very consistently scored close to three. We encode this information in the prior for β_0 which describes the average score of the professors. As in the previous cases, we use a normal distribution centered at three but now we will use a smaller standard deviation of only 0.1 in order to describe that we are relatively confident in values close to three based on the results from the pilot study: $\beta_0 \sim N\ (3, 0.1)$. For the other parameters, we choose priors as in the previous cases. Equation (5) shows the full statistical model.

$$R_i \sim N(\mu_i, \sigma_i) \qquad (5)$$

$$\mu_i \; = \; \beta_0 + \; \beta_1 \times \; graduate\ student_i$$
$$\sigma_i \; = \; \beta_2 + \; \beta_3 \times \; graduate\ student_i$$
$$\beta_0 \sim N(3, 0.1)$$
$$\beta_1 \sim N(0, 0.5)$$

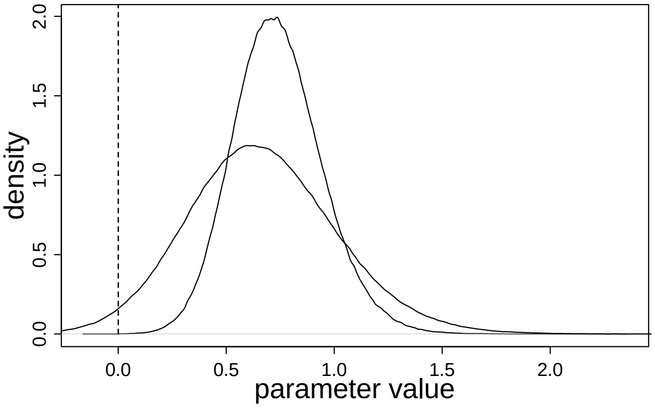
$$\beta_2 \sim Uniform(0,1)$$

$$\beta_3 \sim N(0,0.5)$$

We estimate the model and calculate the posterior of *d* for the R dimension. The posterior distribution (Figure 7) tells us that there is a 95% probability that the difference between the groups ranges between a Cohen's *d* of 0.35 and 1.16 with the highest probability being 0.74.

This result is very different from the one we obtained in our *p*-value based analysis. By incorporating our prior information about the professors into the statistical model, we found evidence that suggests that there really is a medium to strong difference between graduate students and professors with respect to the R dimension. Priors that strongly reflect prior available information are often called informative priors (Gelman et al., 2014) and it is through informative priors that Bayesian data analysis is often better prepared to handle problems with small samples (McNeish, 2016; van de Schoot et al., 2014). However, for transparencies sake, it is important to clearly state the assumptions that went into the formulation of the priors and when using informative priors, we should also conduct a sensitivity analysis in which we re-run the model with an only weakly informative prior and compare the results. Figure 8 shows the posterior that resulted from estimating the model with a weakly informative prior and the posterior that resulted from estimating the model with the informative prior in one panel. The posterior estimated with a weakly informative prior is notably wider than the posterior estimated with an informative prior, reflecting the reduced precision of the estimate. The posterior describes that there is a 95% probability that the difference between the groups ranges between a Cohen's *d* of 0.00 and 1.36 with the highest probability being 0.67. Thus, this posterior is not as strong evidence for a difference between the groups as the one derived from an informative prior but it points in a generally similar direction, thus supporting the general rationale of our informative prior. If the weakly informative prior had yielded totally different results, we would have to question the choice of our informative prior or at least would have to discuss it more thoroughly. As the two posteriors in Figure 8

Figure 8. Posterior density plot of Cohen's *d* for the difference between graduate students and professors on the R dimension. The distribution with a sharp peak resulted from estimating the model with informative priors, the wide distribution resulted from estimating the model with a weakly informative prior. Gray area marks 95% probability interval

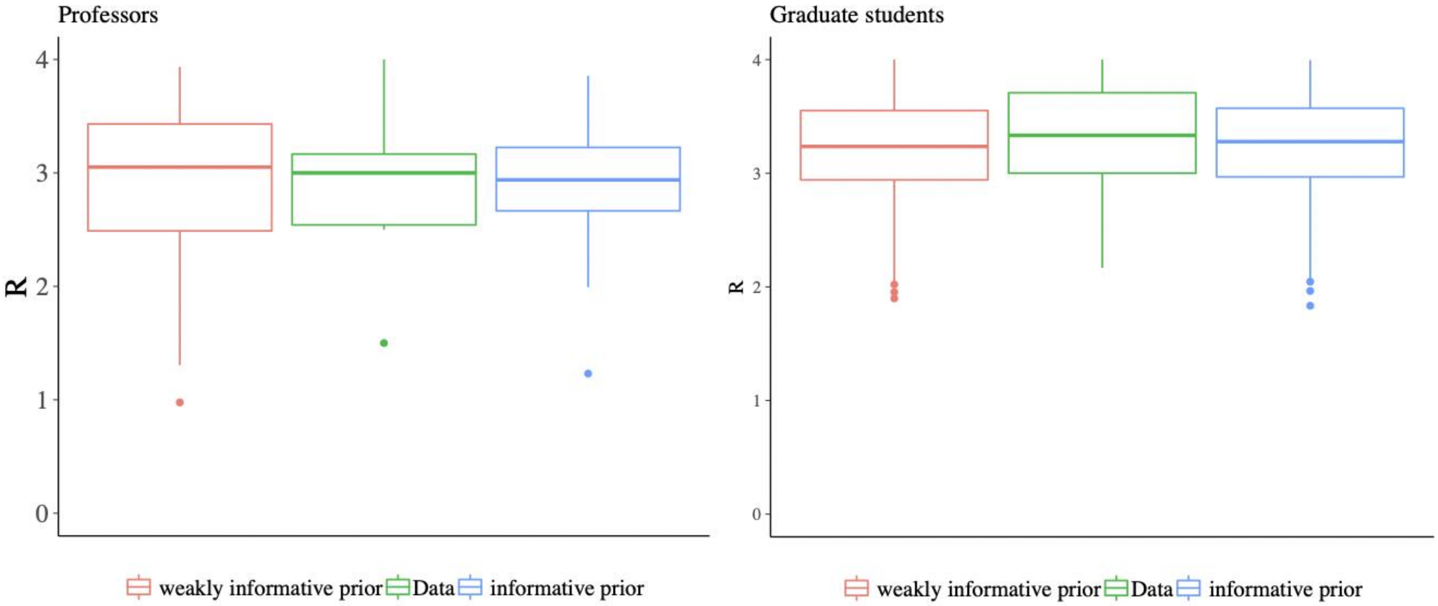


basically represent the two extreme cases of prior specification (weakly informative and strongly informative), we also get a sense of how our prior information influences the result. If we re-estimated the model with another less strongly informative prior, the posterior would be somewhere between the two depicted in Figure 7. Another way to check the accuracy of our model is to look at a posterior predictive plot.

In such a plot, we compare the actual data with the prediction simulated from the fitted model. Figure 9 shows the data from the professors and graduate students, simulated data from the model with weakly informative priors, and informative priors. Comparing the boxplots in the left panel in Figure 9, we see that the simulated results from the model with informative priors have less spread than the simulated results from the model with weakly informative priors which reflects that our informative prior was specified to represent greater confidence in parameter values close to three for the professors. We see a similar thing in the right panel with the graduate students. The simulated data from the model with the informative prior appear to capture the observed data better than the simulated data from the model with weakly informative priors. Thus, the posterior predictive plot lends credibility to how we specified the informative prior and is generally a helpful tool to check model fit.

In sum, the Bayesian analysis allowed us to easily incorporate prior information into our statistical model, which is rather complicated in standard frequentist approaches. This, in turn, allowed us to obtain informative posterior distributions that provide more conclusive information than *p*-values about difference between graduate students and professors on the R dimension.

Figure 9. Posterior predictive check graph showing boxplots of the actual data of the professors and graduate students on the R dimension, and results simulated from the fitted model with weakly informative and informative priors



General Discussion

Making Inferences and Communicating Results

We used a standard *p*-value driven approach and Bayesian data analysis to analyze data in order to answer the question to what extent high school students, graduate students, and professors have different ideas about what scientists do. In case of two clearly different groups (high school and graduate students on the R dimension) and large samples, *p*-values gave us confidence that the observed difference was not random. In case of two relatively similar groups (high school and graduate students on the S and C dimension), standard *p*-value based methods allowed us only to say that there was no evidence for difference between the groups, but not to quantify how good the evidence was for no difference between the groups. In case of a small sample of professors, a large effect size, but a *p*-value only approaching the level of statistical significance, we found ourselves in an inconclusive situation as we had no way of incorporating prior information in the estimation process of the statistical model. Using Bayesian data analysis, our results mirrored those of the *p*-value based approach in case of a clear difference in the data and a large sample. In general, differences between Bayesian and frequentist approach will be small if we have large samples and little prior information. However, in the other two cases, Bayesian data analysis allowed us to reach conclusions that went beyond the *p*-value based analysis. The posterior as the standard Bayesian "result" of an analysis allowed us to easily quantify how good the evidence was for no difference between high school and graduate students on the C dimension. Further, it allowed us to compare that to the difference between high school and graduate students on the S dimension which revealed that in case of the S dimension, the evidence was far from suggesting no difference between the groups. Lastly, when we faced the small sample of professors, we were able to include prior information about the professors into our statistical model, which resulted in a more precise estimation and thus allowed us to conclude that there really is a difference between professors and graduate students on the R dimension. Note that this inclusion of prior information in a prior is distinctly different from including e.g. a covariate in a regression model in order to account for a possible confounding variable. In the latter case, we add variables to better reflect the structure of our design or theoretical model, however, we do not specify prior information about that variable itself.

In general, our results mirror the practical advantages of Bayesian data analysis we find in the literature: 1) Bayesian methods can improve estimation in small sample situation through incorporation of prior information which is often available in science education (McElreath, 2016b; McNeish, 2016; van de Schoot et al., 2014). 2) Once acquired, the Bayesian tool set provides a coherent framework for a range of tasks that often require specialized software packages in the frequentist approach (equivalence testing, penalization, missing value imputation, multi-level modelling). Further, *p*-values are often misinterpreted and lend themselves to dichotomous thinking (Aczel et al., 2017; Cohen, 1994; Gigerenzer et al., 2004; Kahneman, 2012; Kruschke & Liddell, 2017; McShane & Gal, 2017). Posterior distributions as a primary way to communicate results appear to be less prone to dichotomous misinterpretation as their shape describes how confident we are that the true value lies inside a given range. Further, since we have to justify our priors in a Bayesian approach, we make the statistical model explicit and can no longer hide behind jargon such as *t*-test. This should help researchers understand the models of their fellow colleagues better and see the assumptions that the researchers made when analyzing and interpreting their data. An issue which can hardly be overestimated in importance in light of the replication crisis in psychological and social science.

Limitations

The downside of the possibilities that specifying priors provide, is, that it is not always trivial to specify one and that the influence of priors on results has to be discussed as part of a sensitivity analysis (Gelman et al., 2014). Further, fitting Bayesian models is computationally intensive and thus takes longer. The bright side, however, is that Bayesian models may fit where frequentist maximum likelihood methods do not converge (McNeish, 2016; van de Schoot et al., 2014). Thus, apart from large-scale applications where the most prominent practical advantages of the Bayesian approach do not apply because the influence of priors is negligible and the benefit of distributional information ceases to exist in light of very precise estimation, Bayesian models should be considered as a viable alternative. Lastly, there is an issue with the communication of Bayesian results. Standards and conventions of what needs to be communicated in a scientific paper are simply not as developed as for the classical frequentist methods and people are often not as familiar with Bayesian methods as they are with frequentist ones. Thus, more explanation may be needed in order to adequately and successfully communicate the results of Bayesian data analysis. Lastly, we want to clarify that the results of Bayesian data analysis as well as any other statistical approach have to be carefully considered in the context of the actual study in which the data were collected and that the advantages of Bayesian data analysis we have demonstrated cannot solve flaws in the design of a study.

Conclusions

Our examples have focused on two practical advantages of Bayesian data analysis: 1) posterior distributions are more informative than *p*-values which allows for deeper analysis, and 2) incorporating prior information allows for better estimation and thus inferences. We also value the Bayesian approach because it appears to bridge a fundamental gap between data analysis and the knowledge base we build in science education research. Based on substantive theory, we can often make strong assumption about the relationship of variables before we have measured them. This reflects the body of knowledge science education research has built. However, if we follow the frequentist paradigm in our statistical models, we struggle to reflect that knowledge in our statistical models. The prior in Bayesian data analysis allows reflecting our knowledge as researchers in our statistical models. Thus, Bayesian data analysis can help produce statistical models that are closer tied to theory and thus better reflect what we know about the world (Fiedler, 2017; Muthén & Asparouhov, 2012).

References

Aczel, B., Palfi, B., Szollosi, A., Kovacs, M., Barnabas, S., Szecsi, P., Zrubka, M., Gronau, Q., van den Bergh, D., & Wagenmakers, E.-J. (2017). *Quantifying Support for the Null Hypothesis in Psychology: An Empirical Investigation*. PsyArXiv. https://doi.org/10.17605/OSF.IO/ZQKYT

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., … Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z

Brinkworth, R., McCann, B., Matthews, C., & Nordström, K. (2009). First year expectations and experiences: Student and teacher perspectives. *Higher Education*, *58*(2), 157–173. https://doi.org/10.1007/s10734-008-9188-3

Carney, D. R., Cuddy, A. J., & Yap, A. J. (2010). Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychological science*, *21*(10), 1363–1368.

Carpenter, B., Gelman, A., Hoffman, M. D., 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). https://doi.org/10.18637/jss.v076.i01

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist*, *49*(12), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Cumming, G. (2014). The New Statistics: Why and How. *Psychological Science*, *25*(1), 7–29. https://doi.org/10.1177/0956797613504966

De Finetti, B. (1992). Foresight: Its logical laws, its subjective sources. In *Breakthroughs in statistics* (S. 134–174). Springer.

Dierks, P. O., Höffler, T., & Parchmann, I. (2014). Interesse von Jugendlichen an Naturwissenschaften. *CHEMKON*, *21*(3), 111–116. https://doi.org/10.1002/ckon.201410215

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*(3), 193–242. https://doi.org/10.1037/h0044139

Fiedler, K. (2017). What Constitutes Strong Psychological Science? The (Neglected) Role of Diagnosticity and A Priori Theorizing. *Perspectives on Psychological Science*, *12*(1), 46–61. https://doi.org/10.1177/1745691616654458

Gelman, A. (2013). P Values and Statistical Practice: *Epidemiology*, *24*(1), 69–72. https://doi.org/10.1097/EDE.0b013e31827886f7

Gelman, A., & Carlin, J. (2014). Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science*, *9*(6), 641–651. https://doi.org/10.1177/1745691614551642

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis*. http://public.eblib.com/choice/publicfullrecord.aspx?p=1438153

Gelman, A., Hill, J., & Yajima, M. (2012). Why We (Usually) Don't Have to Worry About Multiple Comparisons. *Journal of Research on Educational Effectiveness*, *5*(2), 189–211. https://doi.org/10.1080/19345747.2011.618213

Gelman, A., & Stern, H. (2006). The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant. *The American Statistician*, *60*(4), 328–331. https://doi.org/10.1198/000313006X152649

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The Null Ritual: What You Always Wanted to Know About Significance Testing but Were Afraid to Ask. In D. Kaplan, *The SAGE Handbook of Quantitative Methodology for the Social Sciences* (S. 392–409). SAGE Publications, Inc. https://doi.org/10.4135/9781412986311.n21

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Kahneman, D. (2012). *Thinking, fast and slow*. Penguin Books.

Kaplan, D. (2014). *Bayesian statistics for the social sciences*. The Guilford Press.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573–603. https://doi.org/10.1037/a0029146

Kruschke, J. K., & Liddell, T. M. (2017). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-016-1221-4

Lakens, Daniël. (2017). Equivalence Tests: A Practical Primer for *t* Tests, Correlations, and Meta-Analyses. *Social Psychological and Personality Science*, *8*(4), 355–362. https://doi.org/10.1177/1948550617697177

Lakens, Daniel, Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., Buchanan, E. M., Caldwell, A. R., Van Calster, B., Carlsson, R., Chen, S.-C., Chung, B., Colling, L. J., Collins, G. S., Crook, Z., … Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour*, *2*(3), 168–171. https://doi.org/10.1038/s41562-018-0311-x

McElreath, R. (2016a). *rethinking: Statistical Rethinking book package*.

McElreath, R. (2016b). *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC Press/Taylor & Francis Group.

McNeish, D. (2016). On Using Bayesian Methods to Address Small Sample Problems. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(5), 750–773. https://doi.org/10.1080/10705511.2016.1186549

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2017). Abandon Statistical Significance. *ArXiv e-prints*.

McShane, B. B., & Gal, D. (2017). Statistical Significance and the Dichotomization of Evidence. *Journal of the American Statistical Association*, *112*(519), 885–895. https://doi.org/10.1080/01621459.2017.1289846

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of science*, *34*(2), 103–115.

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological*

Methods, *17*(3), 313–335. https://doi.org/10.1037/a0026802

Nuzzo, R. (2014). Scientific method: Statistical errors. *Nature*, *506*(7487), 150–152. https://doi.org/10.1038/506150a

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716–aac4716. https://doi.org/10.1126/science.aac4716

R Development Core Team. (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. http://www.R-project.org

Sharkawy, A. (2012). Exploring the potential of using stories about diverse scientists and reflective activities to enrich primary students' images of scientists and scientific work. *Cultural Studies of Science Education*, *7*(2), 307–340. https://doi.org/10.1007/s11422-012-9386-2

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), 160384. https://doi.org/10.1098/rsos.160384

Solomon, J., Duveen, J., & Scott, L. (1994). Pupils' images of scientific epistemology. *International Journal of Science Education*, *16*(3), 361–373. https://doi.org/10.1080/0950069940160309

Stamer, I., Kubsch, M., Steiner, M., Höffler, T., Schwarzer, S., & Parchmann, I. (2019). Scientists, Their Work, and how Others Perceive Them: Self-Perceptions of Scientists and Students' Stereotypes. *Research in Subject-matter Teaching and Learning (RISTAL)*, *2*, 85–101. https://doi.org/10.23770/rt1826

van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & van Aken, M. A. G. (2014). A Gentle Introduction to Bayesian Analysis: Applications to Developmental Research. *Child Development*, *85*(3), 842–860. https://doi.org/10.1111/cdev.12169

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*(1), 35–57. https://doi.org/10.3758/s13423-017-1343-3

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's Statement on *p* -Values: Context, Process, and Purpose. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Wentorf, W., Höffler, T. N., & Parchmann, I. (2015). Schülerkonzepte über das Tätigkeitsspektrum von Naturwissenschaftlerinnen und Naturwissenschaftlern: Vorstellungen, korrespondierende Interessen und Selbstwirksamkeitserwartungen. *Zeitschrift für Didaktik der Naturwissenschaften*, *21*(1), 207–222. https://doi.org/10.1007/s40573-015-0035-7

Wetzels, R., & Wagenmakers, E.-J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review*, *19*(6), 1057–1064. https://doi.org/10.3758/s13423-012-0295-x

Citation:

Kubsch, M., Stamer, I., Steiner, M., Neumann, K., & Parchmann, I. (2021). Beyond p-values: Using Bayesian Data Analysis in Science Education Research. *Practical Assessment, Research & Evaluation*, 26(4). Available online: https://scholarwoarks.umass.edu/pare/vol26/iss1/4/

Corresponding Author

Marcus Kubsch
IPN - Leibniz Institute for Science and Mathematics Education
University of Kiel
Kiel, Germany

Email: kubsch [at] leibniz-ipn.de