# The Bayesian Revolution in L2 Research: An Applied Approach


## Abstract

Frequentist methods have long-since dominated in quantitative L2 research (Author, xxxx). Recently, however, a number of fields have begun to embrace alternatives known as Bayesian methods (e.g., Kruschke, Aguinis, & Joo, 2012). Using an open-source approach, this article provides an applied, non-technical rationale for Bayesian methods in L2 research. Specifically, we take three steps to achieve our goal. First, we compare the conceptual underpinning of Bayesian and Frequentist methods. Second, using real as well as carefully simulated examples, we introduce and apply Bayesian methods to various research designs. Third, to promote the use of Bayesian methods in L2 research, we introduce a free, web-accessed, point-and-click software package (https://izeh.shinyapps.io/iiii/) as well as a suite of flexible R functions. Additionally, we demonstrate Bayesian methods for secondary analysis. Practical and theoretical dimensions of a "Bayesian revolution" for L2 research are discussed.


**Keywords:** Bayesian methods, Frequentist methods, effect size, L2 research, research methods

## Introduction

Recent years have seen repeated calls to reform the conventional data analysis practices in the social and behavioral sciences (e.g., Author & Author, xxxx-b; Dienes & Mclatchie, 2017; Etz & Vandekerckhove, in press; Kruschke & Liddell, 2017; Morey, Romeijn, & Rouder, 2016). Most prominent among these calls, however, has been one to shift emphasis away from *Frequentist* methods to *Bayesian* methods. Three critical ingredients are required for such a shift to take place in L2 research. First, in order to embrace the Bayesian methods, we would need to address the difference between the conventional, Frequentist methods and the Bayesian methods. Second, Bayesian methods are to be adapted to be used with a commonly employed set of designs in L2 research (e.g., t-test and correlational designs). Third, and as a very practical matter, software packages that handle Bayesian analyses must be available to a wide audience of

users. It is the goal of this article to address these three key issues and, in doing so, to encourage and enable the use of Bayesian methods in L2 research. All the discussions are accompanied by informationally-rich visuals, and various demonstrations to establish the critical links needed to understand the basics of Bayesian methods with minimal use of technical terms or mathematical expressions. Additionally, following an open-science approach, all the tools, data, and scripts to reproduce the visuals and replicate the analyses are made publicly available to the reader.

## Frequentist and Bayesian Methods: An Introduction

To appreciate the difference between the Frequentist and the Bayesian methods, it is best to apply these methods to a simple research problem. Suppose a researcher administers a single-item survey to determine the *real proportion* of language minority families in a state with a large population of English Language Learners (ELLs) that prefer *bilingual* education (*B*) over *monolingual* education (*M*) for their children (e.g., Bedore, Peña, Joyner, & Macken, 2011; Farruggio, 2010; Ramos, 2007). In this case, parents' preference for the "bilingual" or the "monolingual" (i.e., English-only) education indicates the binary nature of the data that is sought. Given the available resources, the responses from 100 randomly selected parents are collected, 55 of whom prefer "*B*". Thus, the *obtained proportion* of the parents that prefer the bilingual education in this sample is 55%. By contrast, 45% of the parents prefer "*M*" for their children. Also, the 95% confidence interval values for the obtained proportion (i.e., 55%) of parents preferring the "*B*" are: [44.72%, 64.96%]. Figure 1 shows the proportion of preferences for the bilingual education (*B*) as each parent in the sample ($n$ = 100) responds (i.e., "*B*" or "*M*") to the survey question (to explore Figure 1 see https://github.com/izeh/i/blob/master/1.r).

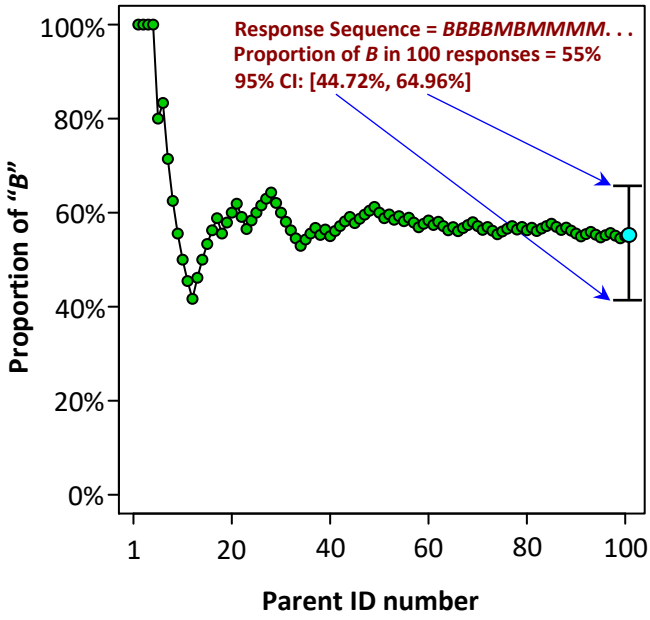


*Figure 1*. Proportion of preferences for bilingual education. *"B"* denotes preference for bilingual education and *"M"* denotes preference for monolingual education.

At this point, the critical question is: Given that we have data from only 100 parents in the state, can we discover the *real* proportion of preferences for bilingual education *in the entire state*? This question has a Frequentist as well as a Bayesian answer.

From the Frequentist perspective, the answer to this question relies on the Frequentist theory. According to this theory, there is surely one objective answer to the question above. However, there will always be uncertainty in any one answer (i.e., point estimate; here 55%) obtained from any one study with a limited sample size (e.g., 100 parents). To incorporate this uncertainty in any obtained answer, Frequentists use a confidence interval (CI) whose interpretation requires close attention. For example, the 95% Frequentist confidence interval of [44.72%, 64.96%] obtained from our above survey "would indicate that over long-run frequencies [i.e., infinitely many repetitions of the survey], 95% of the confidence intervals constructed in this manner (e.g., with the same sample size, etc.) would contain the true

population value" (Depaoli & van de Schoot, 2017, p. 257). To better understand the nature of this interpretation, Figure 2 shows a possible set of results from only 20 such repetitions of our survey (to explore Figure 2 see https://github.com/izeh/i/blob/master/2.r). The filled circles represent the observed proportion of the parents that prefer bilingual education in each of these 20 repetitions of the survey. The solid horizontal lines passing through the filled circles are the 95% confidence intervals for the obtained proportion of preferences for "*B*" in each of these 20 repetitions of the bilingual education survey.

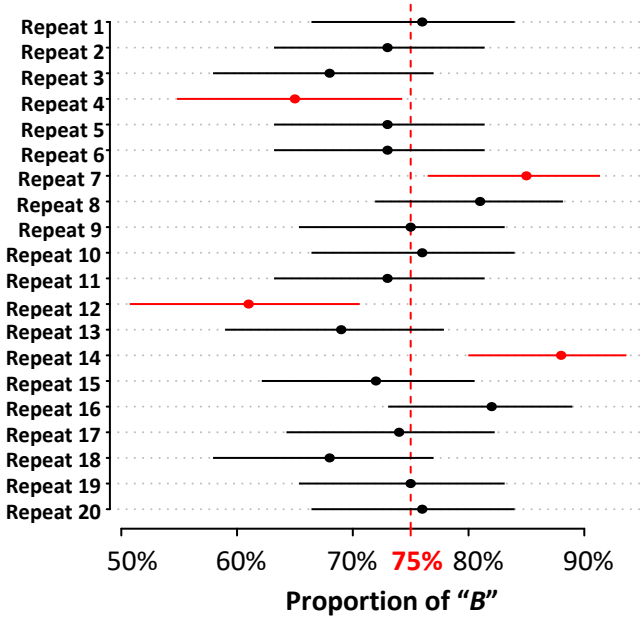


*Figure 2.* Twenty repetitions of the same bilingual education survey. The vertical red line represents the real (i.e., state-wide) proportion of preferences for bilingual education.

Let us assume for the sake of this demonstration that the real proportion of preferences for bilingual education in the population of parents is 75% (as shown by the vertical dashed red line in Figure 2). In this case, some of the obtained proportions (filled circles) in these 20 repetitions have either egregiously underestimated or overestimated the real proportion of preferences for bilingual education. These observed proportions are indicated in red as are their associated 95% confidence intervals, which do not contain the real proportion of preferences for bilingual

education (i.e., 75%). Of course, 20 repetitions are not infinitely many repetitions. In theory, if repeated infinitely many times, 95% of the obtained confidence intervals will contain the real proportion of preferences for "*B*" that our researcher is interested in. Based on this perspective, therefore, the Frequentist answer to the *critical question* relies on a procedure that in the long run produces a specified rate of being correct (e.g., 95%), and a specified error rate (e.g., 5%). Consequently, the so-called 95% confidence level often attached to an obtained CI in reality applies to a Frequentist, long-run procedure in which infinitely many intervals are assumed; it does not denote that interval estimates obtained from a single study (here 44.72% - 64.96%) have captured the population value with 95% certainty (see Thompson, 2006, p. 204).

From the Bayesian perspective, however, this long-run procedure and the subsequent interpretation is considered unnecessarily complex. That is, such a Frequentist interpretation not only is not desired, but it also could be a source of confusion for a researcher wanting to interpret her/his single study's obtained results. Surely, what one seeks to have is X% certainty that a single obtained interval from her/his study has captured the population value.

The Bayesian methods do not require thinking in Frequentist terms. Rather, they start from the position that when a parameter is unknown (e.g., proportion of parents preferring "*B*"), then it is wiser to think of that parameter as being a variable (rather than thinking of that parameter as having a single true value as in the Frequentist approach) with a full range of possibilities governing its magnitude. As one of the ways to apply this view to our bilingual survey from above, the Bayesian method might begin by asking our researcher to use the prior empirical findings relevant to the phenomenon under study, and/or the theoretically defensible expectations for the phenomenon under study to define an expected range for the *real* proportion of the preferences for "*B*" prior to conducting the survey. Given such knowledge, some of the

values in this expected range may be more strongly expected and some less. The resultant expected range along with the weights given to the individual values in it lead to the formation of a "prior" distribution. For example, a review of the past literature might reveal that (a) the proportion of language minority parents that prefer bilingual education has been varying between 60% and 80% in the state of interest, and (b) efforts and investments in promoting bilingual education in that state have been constantly increasing over time. Based on this knowledge, the values of proportion found to be smaller than 60% or larger than 80%, although possible, are logically less likely to represent the real proportion of preferences for "*B*" in the population of parents. Figure 3 shows a possible prior distribution (see next section for prior appropriateness checking) that would match the researcher's expectations described above (to explore Figure 3 see https://github.com/izeh/i/blob/master/3.r). Displayed for better visualization, the upward-pointing arrows in the middle denote the higher weights given to individual values between 60% and 80%. By contrast, the downward-pointing arrows denote the successively lower weights assigned to individual values outside 60% and 80%. Such a weighting scheme often results in prior distributions that resemble a bell-curve of some kind peaked over the expected range (here 60% - 80%).

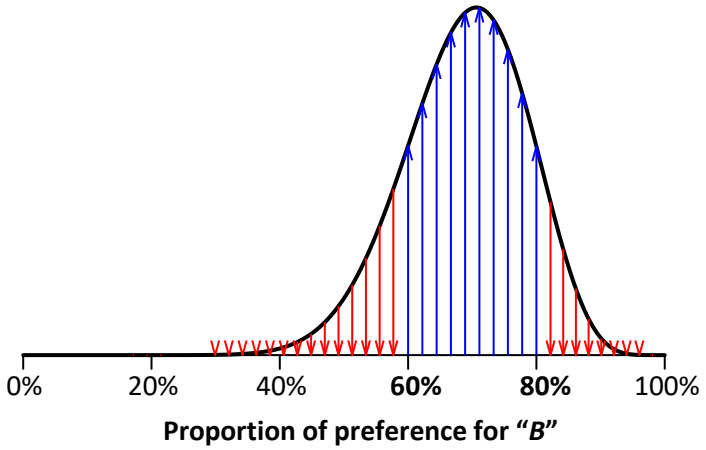


*Figure 3.* Prior distribution for the proportion of preference for bilingual education.

Now that the prior distribution is at hand, the next step is to obtain the likelihood function for the obtained proportion of preferences for "*B*". The likelihood function is easy to obtain. Because, depending on the nature of the study data, the likelihood functions are either well known or relatively easy to construct. In our case, because the nature of the survey data is binary (i.e., "*B*" or "*M*"), the likelihood function is known to be a "Binomial" one. All we need to do is to input the number of parents who preferred "*B*" (i.e., 55), and the total number of parents surveyed (i.e., 100) to a Binomial formula, and indicate the place for the unknown proportion of preferences for "*B*" in the formula by an "x", perhaps using a software package (see https://github.com/izeh/i/blob/master/4.r for an R implementation). Figure 4 shows the likelihood function for our example (to explore Figure 4 see https://github.com/izeh/i/blob/master/5.r).

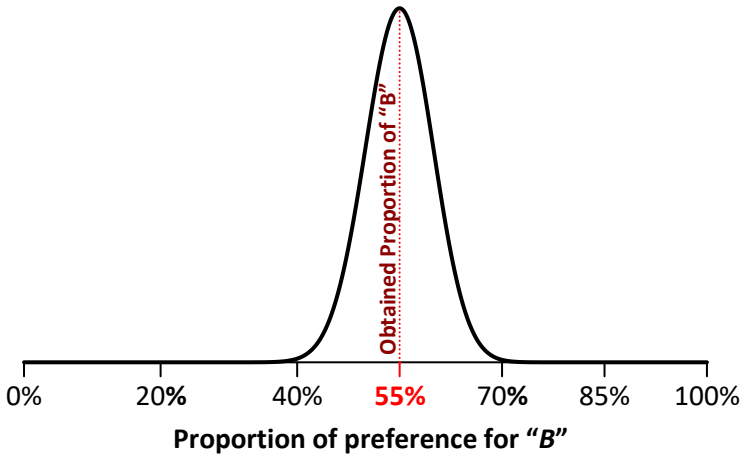


*Figure 4.* Likelihood function for the proportion of preference for bilingual education.

In terms of weighting, the likelihood function automatically assigns the highest weight to the obtained proportion of "*B*" (i.e., 55%). This is almost always the case, because, as implied earlier, likelihood functions are formulas that operate solely on the basis of the obtained data. Thus, they recognize the obtained proportion of preferences for "*B*" as the most likely estimate of the real proportion of preferences for "*B*" in the population of parents and all other possible

estimates further away from this estimate as successively less and less likely. Now that we have the two essential ingredients of a Bayesian approach (i.e., prior and likelihood), it is time for the Bayesian mantra (i.e., Bayes' theorem):

$$\textit{Prior} \times \textit{Likelihood} \propto \textit{Bayesian Result} \quad (1)$$

where "$\propto$" (is proportional to) denotes the fact that a Bayesian result from this equation remains proportional to its proper form until scaled by a normalizing constant (see Gelman, Carlin, Stern, & Rubin, 2014). For simplicity's sake, the reader may take "$\propto$" as "=". Equation 1 is the main guiding equation in Bayesian methods applied to ANY research problem. And the Bayesian result obtained is the only result that an expert researcher will need to describe and interpret. At no point will one need to refer to the infinitely many repetitions (i.e., long-run frequencies) of the exact same survey necessary under the Frequentist paradigm. Essential to know is that the Bayesian result is better known as the "*Posterior*". Per our Bayesian mantra, the posterior is obtained by multiplying the prior distribution by the likelihood function. Figure 5 illustrates this multiplication to obtain the posterior for our example (to explore Figure 5 see https://github.com/izeh/i/blob/master/6.r).

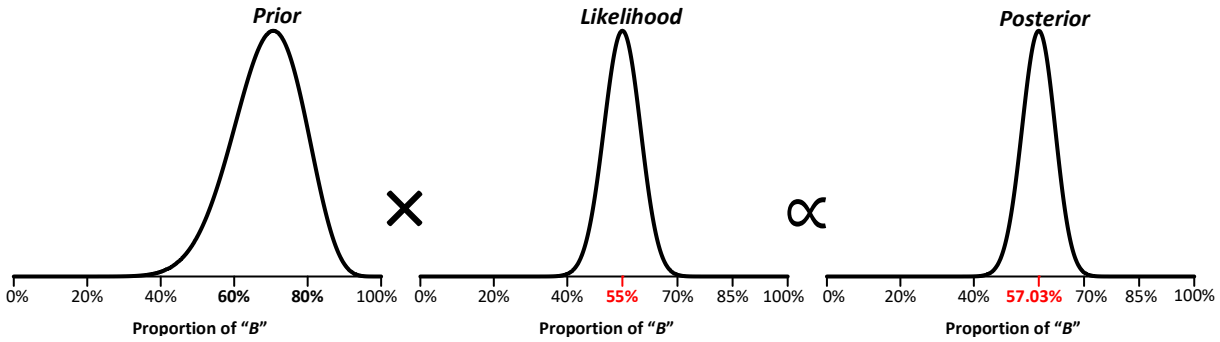


*Figure 5.* Steps to obtaining the Bayesian result (i.e., posterior) for estimating the proportion of preferences for bilingual education.

At this point, we can more precisely concentrate on our obtained posterior. Figure 6 shows the posterior for our example with more details added to it to help the accurate interpretation of our Bayesian result (to explore Figure 6 see https://github.com/izeh/i/blob/master/7.r). As is discussed next, these details provide direct insights into finding out what the real proportion of parents' preferences for "*B*" in the entire state (population) could be.

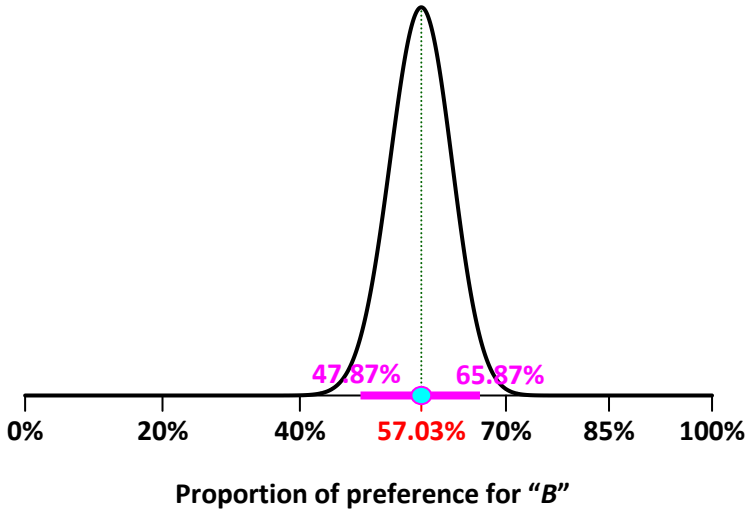


*Figure 6.* Posterior distribution for the proportion of preference for bilingual education.

The confidence interval-like horizontal line segment at the bottom of Figure 6 covers 95% of the most highly weighted (i.e., densest) areas of the posterior. Proportion values inside this 95% range are more credibly likely to represent the real proportion of preferences for "*B*" than others in the posterior. As such, this confidence interval-like range is sometimes referred to as a "Highest Density Credible Interval" (henceforth *credible interval*). Such a credible interval is quite helpful in describing and interpreting a posterior.

With the help of this credible interval, our researcher is now able to state that there is 95% probability that the *real* proportion of preferences for "*B*" in the population of parents could credibly range between 47.87% and 65.87%. Notice the brevity and the directness with which a single obtained Bayesian credible interval describes the candidate values representing the real

proportion of preferences for "*B*" in the population of parents. Also, note that the values of proportion closer to the center (filled circle) of this credible interval are still more likely to represent the real proportion of preferences for "*B*" in the population of parents than others. As we discussed above, none of these informative properties could be interpreted from a single obtained Frequentist confidence interval.

**Putting priors to the test**

In the previous section, we discussed that a Bayesian method starts by choosing a prior. Often, however, the prior distribution picked for estimating a parameter must pass a test for it to prove plausible. There are several ways of evaluating the plausibility of a prior depending on the nature of the parameter at hand, as well as the type of prior selected. In the case of estimating the proportion of parents supporting bilingual education described in the previous section, we used a type of prior that belonged to the "Beta" family. Beta priors are naturally bound between 0 (or 0%) and 1 (100%). Thus, they could be one possible prior type for estimating a parameter (e.g., proportion, eta squared effect size; see Author & Author [XXXX]) that ranges between 0 (or 0%) and 1 (or 100%). Albert (2009) suggests that a beta prior distribution may be specified "through statements about the percentiles of the distribution" (p. 23). In non-technical terms, even if past research shows that the proportion of language minority parents that prefer "*B*" for their children varies between 60% and 80%, we might not exactly know how well such findings do at representing the true proportions of preference for "*B*" across the state. That said, it would be perhaps unrealistic to think that the degree of representativeness for previous findings could be fairly high or fairly low. If one chooses to express this degree of representativeness in percentages (i.e., from not representative; 0% to completely representative; 100%), then conservatism dictates that a reasonable range for this representativeness could start from mid-low

(e.g., 40%) to mid-high (e.g., 60%). This means we can specify different priors that separately take this range for representativeness (i.e., 40%, . . ., 50%, . . ., 60%) for the past research findings into consideration and then obtain the corresponding posteriors under all such priors. To do this, we suggest using our suite of R functions accessible by running the following in R or RStudio®:

```
source("https://raw.githubusercontent.com/izeh/i/master/i.r")
```

The first step would be to obtain a set of priors (e.g., 10) that incorporate the range of 40% to 60% for the representativeness of past survey findings (i.e., 60% - 80%). The R function "`beta.id`" is designed for this purpose:

```
I = beta.id(Low = "60%", High = "80%", Cover = seq(.4, .6, l = 10))
```

Now, we have 10 different prior specifications each of which incorporating in it a different level of representativeness (i.e., 40% (or .4) - 60% (or .6)) for the past research findings (i.e., 60% - 80%), all stored in "`I`". Each of these 10 prior distributions can be individually inspected using the R function "`prop.bayes`". For example, to see the last (i.e., 10th) prior which was also used in the previous section (see Figure 3) we can use:

```
prop.bayes(a = I[1,10], b = I[2,10], dist.name = "dbeta", show.prior = TRUE)
```

Or to see the first prior displayed below in Figure 7 we can use:

```
prop.bayes(a = I[1,1], b = I[2,1], dist.name = "dbeta", show.prior = TRUE)
```

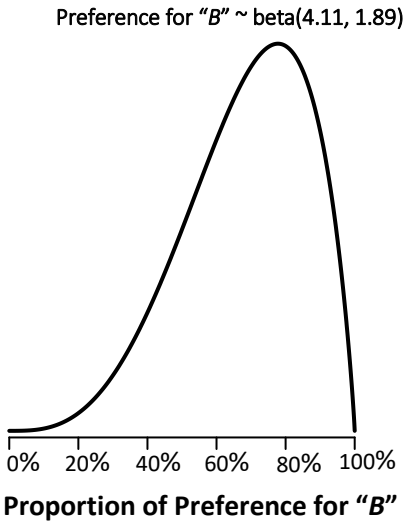


*Figure 7.* The first prior distribution for the preference for "*B*" (see text).

The next step is to obtain the Bayesian result (i.e., posterior) using all these different priors one at a time and compare their resultant 95% credible intervals. Egregious differences among the 95% credible intervals would indicate that our results are sensitive to uncertainty about the representativeness of past research findings. When such notable differences occur, we have failed the test of robustness under our choices of prior. To perform these analyses and compare their 95% credible intervals, we can once again use the function "`prop.bayes`":

```
prop.bayes(a = I[1,], b = I[2,], dist.name = "dbeta", scale = .1, top = 1.055)
```

The Bayesian posteriors along with their 95% credible intervals are provided in Figure 8.

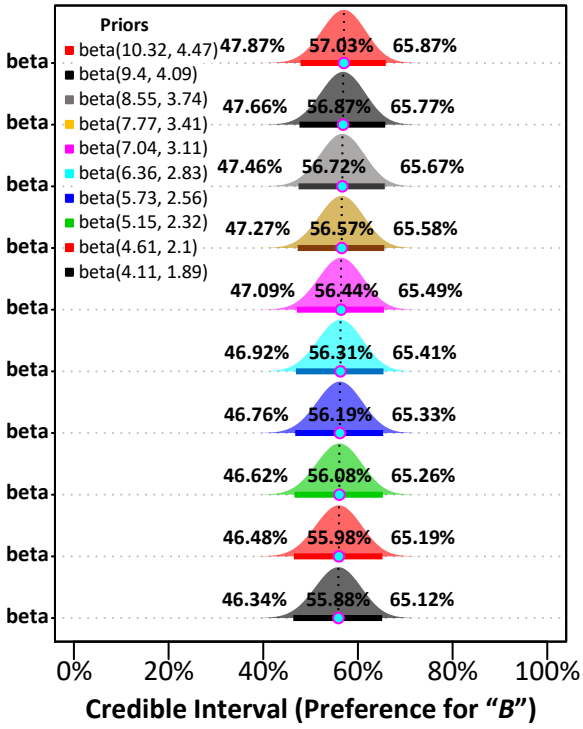


*Figure 8.* Bayesian posterior credible intervals under various Beta priors

As can be seen, although the priors are different, the posteriors are fairly aligned with each other with no egregious differences among their 95% credible intervals. After taking a reasonable set of candidate priors, the visual inspection of the credible intervals is critical in demonstrating the robustness of results under the choice of priors. In the following sections, we will see that in various situations, the nature of the parameter at hand and the type of common priors employed to describe it allow us to conduct other forms of robustness analyses.

*Skepticism and lack of prior knowledge*

In some cases, prior knowledge is absent, diminished, or its credibility might be under question. In such situations, priors that concentrate their weight on (i.e., are peaked over) a certain range for a parameter may be easily prone to biasing a Bayesian result (i.e., posterior). Defining a prior distribution that expresses the state of neutrality or a lack of knowledge is one way to avoid such potential biases. Several seminal works have looked at this issue from

perspectives that require both space and technical background knowledge (e.g., *information theory*; Jaynes, 2003; *invariance to transformation*; Jeffreys, 1961; *contribution of prior measured in datapoints*; Liang, Paulo, Molina, Clyde, & Berger, 2008). In this introductory discussion, however, we tend to simply refer to priors that express a lack of or minimal prior knowledge as "broad" or "minimally informative". As we shall see, reasonableness must always play a role in defining such priors depending on the nature of the parameter at hand and the type of prior meant to be used with it.

Let us use an actual example in which lack of prior knowledge is best evident. Author & Author (xxxx-b) surveyed the application of two effect size variants, eta-squared ($\eta^2$) and partial eta-squared ($\eta_p^2$), in a sample of 156 studies that used these two effect sizes from various L2 journals published between 2005 and 2015. Surprisingly, the authors found that in 34 studies, the primary authors of the published L2 research had erroneously reported and interpreted partial eta-squared effect size in place of eta-squared effect size (for consequences of this misreporting see Author & Author, xxxx-b). This indicated that 21.79% (i.e., $\frac{34}{156} \times 100$) of the collected sample of L2 studies had misreported these two effect variants. But assuming that L2 researchers did not largely learn how to use and distinguish these two variants of effect size from each other (i.e., independence of observation), the question of interest is: What is the actual proportion of the studies that contain this erroneous reporting, and how prevalent this erroneous reporting is across all L2 studies that report these two measures of effect size? Since this was the first survey of this type in L2 research, no specific prior knowledge in L2 research is available to refer to as a knowledge base. Also, similar studies in sister fields such as psychology (Pierce, Block, & Aguinis, 2004) and communication (Levine & Hullett, 2002) tend to only narratively describe the existence of a confusion in using the two variants of effect size among researchers in their

respective fields without offering much quantifiable evidence. With such highly restricted knowledge base, defining an informative prior distribution may not be possible. What is needed, however, is a "broad" or "minimally informative" prior distribution that assigns almost equal weights to most possible values (i.e., 0% - 100%) representing the proportion of studies that misreport eta- and partial eta-squared as two measures of effect size in L2 research. Many Bayesian analysts have argued that it is always wise to exclude extremely unrealistic values that may not represent the possible magnitude of the parameter (here the proportion of misreported L2 studies) under estimation (e.g., Gelman et al., 2014; Kruschke, 2015; McElreath, 2016). In our case, to assume that the misreporting proportion of the two measures of effect sizes, eta- and partial eta-squared, in L2 research could be close to ~0% or ~100% is unequivocally unrealistic. A broad prior then could be one (a) whose effective weight concentration spans over most possible values for the misreporting proportion excluding the unrealistic ones (e.g., ~0% and ~100%), and thus (b) which is not skewed toward a particular side in the parameter range (i.e., is symmetric slightly pivoting on 50%). One such broad prior is shown in Figure 9. Figure 9 can be easily replicated using our R function "`prop.update`":

```
prop.update(a = 1.2, b = 1.2, show.prior = TRUE, prior.scale = .5, top = 1.6)
```

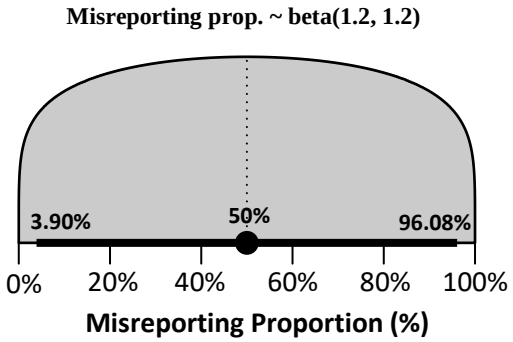


*Figure 9.* A broad prior expressing lack of knowledge for misreporting proportion.

With this broad prior at hand, we can proceed with estimating the proportion of studies that misreport eta-squared ($\eta^2$) and partial eta-squared ($\eta_p^2$), as two measures of effect size, in published L2 research. The function "`prop.update`" can be called again to see how our broad prior knowledge is changed in light of the 34 cases of effect size misreporting out of 156 studies examined by Author & Author (xxxx-b):

```
prop.update(yes = 34, n = 156, a = 1.2, b = 1.2, scale = .2, top = 5, prior.scale =
1.3)
```

The result of our analysis is displayed in Figure 10.

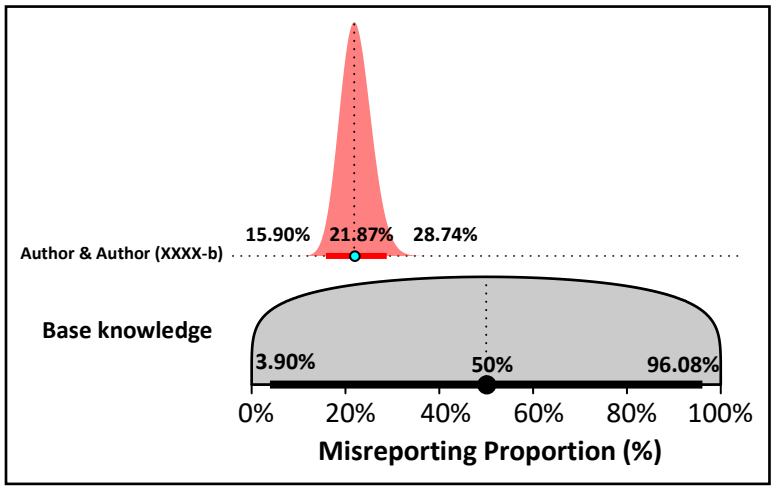


*Figure 10*. Updating a broad knowledge base in light of misreporting cases found in Author & Author (xxxx-b).

Our analysis shows that there is 95% probability that misreported proportion of studies range between 15.90% and 28.74%. Since we used a minimally informative (i.e., nearly flat) prior to express our lack of prior knowledge, these Bayesian results may not need to be checked for their robustness. As a way to get more familiar with other families of priors, however, we can again submit our choice of prior to testing and visually examine our Bayesian results. We can select from a variety of different families of priors in addition to "Beta". These other families are first

positioned such that they, just like our Beta prior, cover the entire range of 0% to 100% for misreporting rate (i.e., our parameter) of the two effect sizes but then cut for any additional coverage for values that do not fall within 0% to 100%. For example, the familiar Normal distribution which is naturally boundless (i.e., goes from –infinity to +infinity) is first positioned so that, like our Beta prior, it reflects neutrality and symmetry (e.g., centered at .5 or 50%) but then cut everywhere except for areas falling between 0 and 1. Here we use two other families of distributions in addition to "Beta", namely "Normal" and "Cauchy" (see next section on effect size). Both of these distributions are naturally symmetric, but we can position them between 0 (or 0%) and 1 (or 100%) while centering them at .5 (or 50%). Note that in R and some other software packages (e.g., JAGS, WinBUGS), distribution names start with a "`d`" (standing for *density*). Examples include "`dnorm`" for Normal, "`dcauchy`" for Cauchy, and "`dbeta`" for Beta distribution. We can use the R function "`prop.bayes`" to test these three prior families all at once:

```
prop.bayes(a = c(1.2, .5, .5), b = c(1.2, 1, 1), dist.name = c("dbeta","dnorm",
"dcauchy"), scale = .075, top = 1.4, yes = 34, n = 156)
```

The resultant posteriors along with their 95% credible intervals are shown in Figure 11.

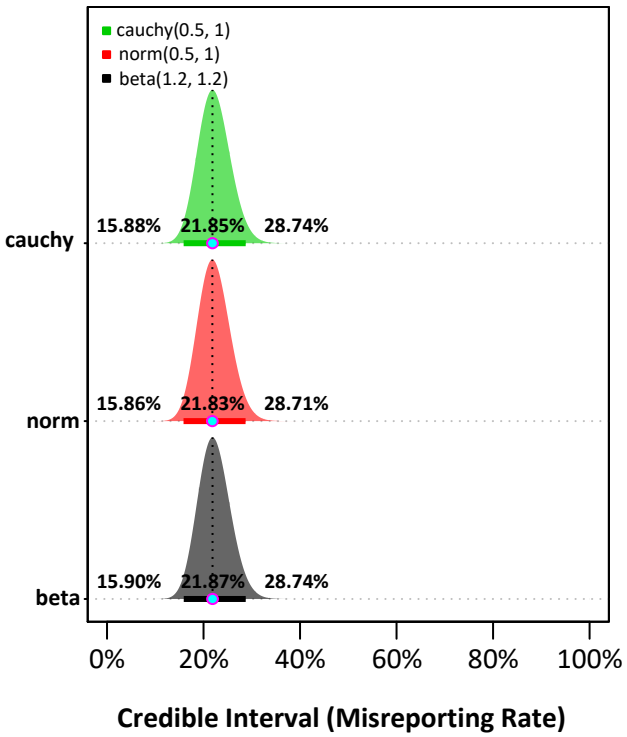


*Figure 11*. Posterior results under different families of priors.

As shown in Figure 11, the results under these three families of priors barely change. Indeed, even if the width (spreadoutness) of Normal and the Cauchy priors are increased by a factor of 10 (i.e., from 1 to 10) no major change in the posteriors occurs:

```
prop.bayes(a = c(1.2, .5, .5), b = c(1.2, 10, 10), dist.name = c("dbeta",
"dnorm","dcauchy"), scale = .075, top = 1.4, yes = 34, n = 156)
```

Figure 12 shows the result of the ten-fold increase in the width of the Normal and Cauchy priors. As noted earlier, the use of sensitivity analysis is not really warranted when a broad prior is used. This demonstration, therefore, serves to visually show that after selecting a broad prior, the results remain invariant to yet wider prior specifications. Thus, it is safe to believe that there is 95% probability that the proportion of studies that misreport the two measures of effect size, eta- and partial eta-squared, in L2 quantitative research ranges between ~15.9% and ~28.7% as indicated by our 95% high-density credible intervals.

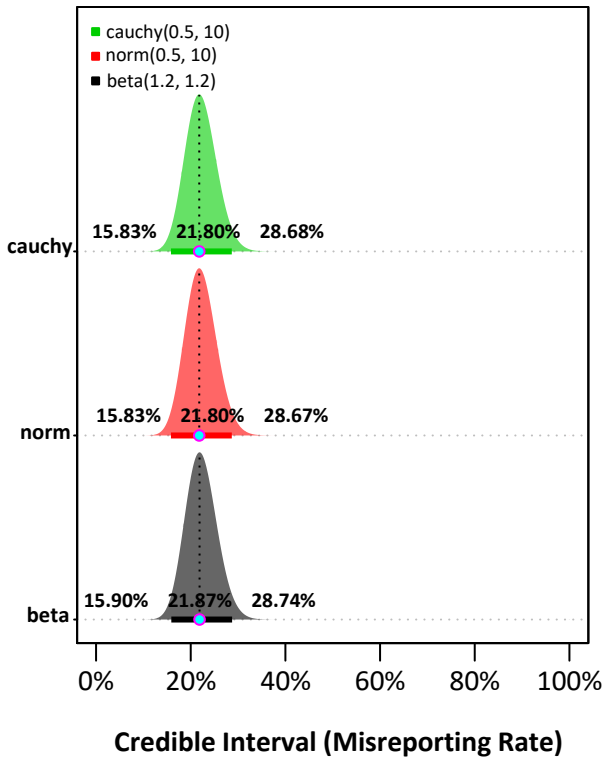


*Figure 12*. The result of a ten-fold increase in the width of the Normal and Cauchy priors.

*Letting priors arise*

Many of us as applied linguists would agree that the knowledge generated from our studies must play a role in informing future replication efforts (see Marsden, Morgan-Short, Thompson, & Abugaber, in press; and Porte, 2012 for rich discussions of replication in L2 research). Bayesian methods are uniquely designed so that each future replication could build on the knowledge generated by any number of replication works conducted before it (see Note). This feature of Bayesian methods is so boundless that it is often said that *yesterday's posterior is today's prior* (see Lindley, 2000). To better see this in action, suppose that two other surveys at two different points in time had targeted the preference of language minority parents for bilingual education before our current survey discussed in the previous section. A Bayesian framework allows us to cumulatively incorporate these two other surveys' results into our

current survey in a step-wise fashion. That is, one can (a) start with a broad knowledge base, (b) use that broad knowledge base as a prior for the first available survey to obtain the posterior, (c) use that posterior as prior for the second survey to obtain a second posterior, and finally (d) use the posterior of the second survey as prior for the current survey, obtain the final posterior, and describe it using 95% credible intervals as the most current result. This step-wise Bayesian updating process is implemented in our R function "`prop.update`". To use the function, suppose the first and oldest survey came from 70 parents, 27 (39%) of whom preferred bilingual education, and the second survey was based on 84 parents, 31 (37%) of whom favored bilingual education for their ELLs (English Language Learners). Recall that our current survey (see Figure 1) showed that 55 out of the 100 parents support bilingual education. Now, a call to function "`prop.update`" can be made to incorporate both of the previous surveys' results in our current replication survey using a broad prior base:

```
prop.update(n = c(70, 84, 100), yes = c(27, 31, 55), a = 1.2, b = 1.2, dist.name =
"dbeta", scale = .086, top = 1.6)
```

The result of this step-wise Bayesian updating is shown in Figure 13. As can be seen, we started from a very broad knowledge base that allowed us to believe almost any proportion (0% - 100%) could be a candidate value for representing the proportion of parents that prefer bilingual education. But then this broad knowledge base was updated by the first survey conducted on the matter. Still, the second survey built on both the initial knowledge base as well as the result of the first survey and this updating went on until the most recent survey was carried out. Other than *letting the priors arise* in the process rather than specifying them in advance, the end result of such updating processes is one final posterior that, founded upon previous replication attempts, will concentrate narrowly on the proportion values (or any other parameter of interest) that represent the parents' view regarding bilingual education. In the later sections, we will return

to this updating process to generate a prior based on the findings of previous replication research and extend it to situations where our parameter of interest is a standardized mean difference effect size (i.e., Cohen's *d*).

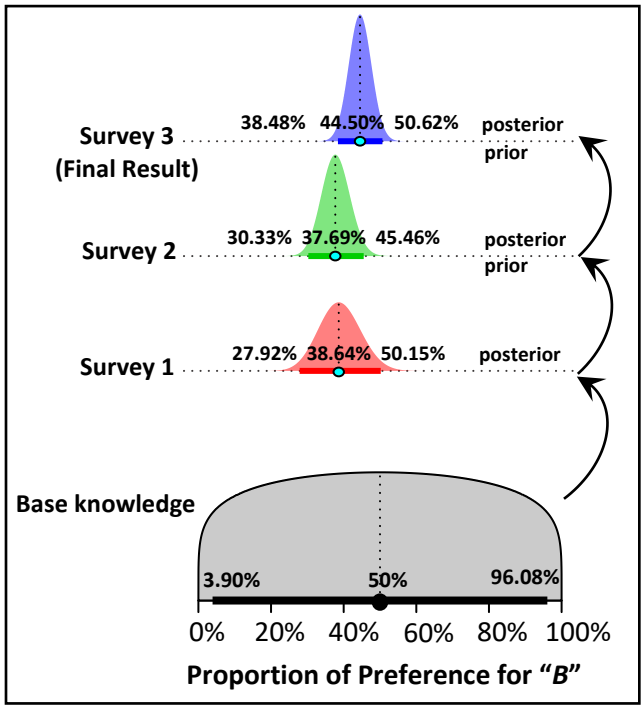


*Figure 13*. Step-wise updating of three bilingual education survey using a broad prior.

In the next section, we present an application of Bayesian methods for one of the most commonly employed statistical analyses in L2 research, the *t-test*, (Author et al, xxxx; Larson-Hall, 2016). Through the Bayesian method, we add a new application to t-tests so that in addition to being used for testing the validity of a null hypothesis, t-tests become vehicles for estimation of the real effect (i.e., effect size) of a treatment. In addition to a repository of highly flexible R functions, we also introduce a free, online, point-and-click software package (https://izeh.shinyapps.io/iiii/) that automates some of the steps involved. As will be shown, this Bayesian application of t-tests can also be used for the Bayesian estimation of the real effect of a

treatment (i.e., effect size) from a previously published study using only the basic information available in that study.

**Bayesian Methods as Applied in t-test Designs in L2 Research**

The Bayesian method discussed in the previous section also applies to designs that use *t-tests*, which are ubiquitous in L2 research (Author, xxxx; Larson-Hall, 2016; Linck & Cunnings, 2015). And the approach that we take to run Bayesian t-tests, an "effect size" approach, concurs in the belief that "the primary product of a research inquiry is one or more measures of effect size, not *p*-values" (Cohen, 1990, p. 1310). To be clear, t-tests are analytic tests that are used to evaluate *if there is an effect* (i.e., null hypothesis testing; *p*-value) for a treatment in pre-post designs (paired-samples t-test), experimental designs with two groups (independent samples t-test), and one-sample designs (one-sample t-test), the last of which is less commonly found in L2 research (see Larson-Hall, 2016, p. 270). The marriage of the Bayesian methods and the effect sizes from such designs allows for estimating the real size of an effect for a treatment from the above-mentioned designs. In our view, this significantly adds to the applicability and utility of t-tests in L2 research.

Let us then apply the Bayesian t-test method to a meaningful L2 research example as we did when discussing proportion-type data in the previous sections. Suppose a researcher is interested in finding out the real effect of an L2 treatment on improving the explicit knowledge (DeKeyser, 2015; Lyster & Sato, 2013) of 60 high-intermediate English as a Foreign Language (EFL) learners with respect to Type III conditionals (e.g., *If I had arrived earlier, I could have caught the bus*). The schematic design of this study is shown in Figure 14.

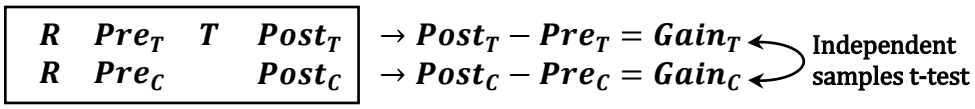


*Figure 14*. Pre-post-control design layout. *R = Random assignment*; *T = Treatment*; *C = Control; Pre = Pre-test; Post = Post-test.*

Based on this *Pre-Post-Control* design, the participants are randomly assigned to either the treatment group ($n$ = 30) or the control group ($n$ = 30) to protect the study outcome from some of the design's internal validity threats, e.g., regression to the mean (see Campbell & Stanley, 1963). Then, following the pre-test and treatment, the researcher administers a posttest to measure the difference in the level of the explicit knowledge of Type III conditionals gained by the two groups. To measure explicit knowledge (see Ellis, 2009), both groups are to complete an untimed error correction test (ECT) consisting of 15 sentences 10 of which contain different number grammatical errors in the use of Type III conditionals. The scoring scheme used for Type III conditionals often involves awarding a combination of half-points and whole-points depending on what feature (e.g., correcting the past modal: *"would / could / . . ."* 1 point, correcting the past participle form: *"caught"* .5 point) in the *conditional* or the *main* clause is appropriately corrected by a participant (see Izumi, Bigelow, Fujiwara, & Fearnow, 1999). In total, 25 points are allowed on the entire error correction test. Recent research (e.g., Shintani, Ellis, & Suzuki, 2014) suggests that it is reasonable to believe that this scoring scheme would result in scores complying with the assumption that such scores belong to normally shaped populations. Finally, it is important to note that there are several ways to analyze the above *Pre-Post-Control* design (see Salkind, 2010). Here we follow a simple approach which requires us "to compute for each group pretest-posttest gain scores and to compute a *t* [i.e., independent samples *t*] between experimental and control groups on these gain scores" (Campbell & Stanley, 1963, p. 23). With these details in mind, let us simulate such a study, and then employ a

Bayesian independent-samples t-test to estimate its possible effect. Figure 15 graphically shows the design, raw gain scores, and the immediate results of this simulated study (to explore Figure 15 see https://github.com/izeh/i/blob/master/11.r).

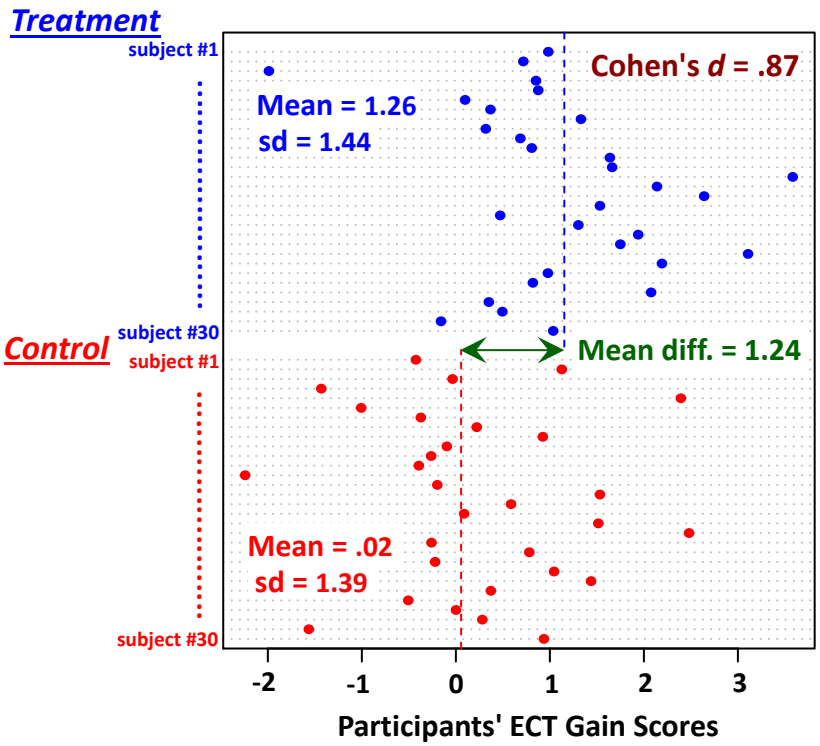


*Figure 15.* The design and raw gain scores (posttest – pre-test) of the participants in the simulated study. ECT = Error Correction Test. Each grey, horizontal, dotted line denotes a participant. The vertical dashed lines denote the mean of each group's gain scores. Mean diff. = difference between the means of groups' gain scores.

Table 1 presents the full descriptive and the conventional Frequentist results (e.g., confidence interval) of this study.

Table 1. *Frequentist Study Results for EFL Learners in the Simulated Study* (*N* = 60)

| | Descriptive (Gain Scores) | | | | Inferential | |
|---|---|---|---|---|---|---|
| Group | *n* | *M* (*SD*) | *ES* (*d*) | 95% $CI_{(d)}$ | *t* (*df*) | *p*-value |
| Treatment | 30 | 1.26 (1.44) | **.87** | [.33, 1.40] | 3.35 (58) | .001 |
| Control | 30 | 0.02 (1.39) | | | | |

*Note. M* = Mean; ES = Effect size; *d* = Cohen's *d*.

The effect size (i.e., *d* = .87) along with its 95% confidence interval (i.e., 95% $CI_{(d)}$ [.33, 1.40]) obtained from our simulated study (Table 1) are both subject to Frequentist interpretations. Recall from our discussion in the previous sections that from the Frequentist perspective these results can be theoretically seen as just one set of possible results from among many more in the long chain of repetitions of the exact same study on Type III conditionals. For example, let us assume that in reality our L2 treatment is able to produce an effect quantified by a Cohen's *d* effect size of .5, then a possible set of results from only 20 repetitions of our exact same study on Type III conditionals is presented in Figure 16 (to explore Figure 16 see https://github.com/izeh/i/blob/master/9.r).

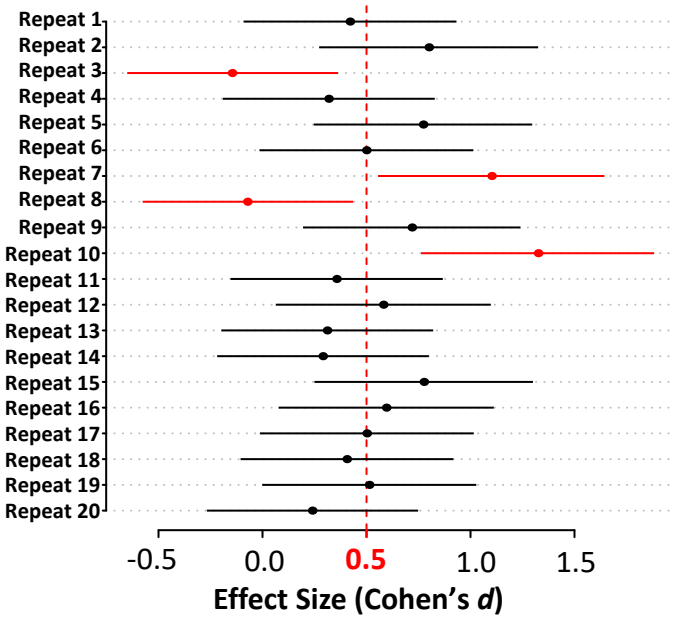


*Figure 16.* Twenty repetitions of the same study on Type III conditionals. The vertical red line represents the real (i.e., population) size of effect for the L2 treatment.

As with the survey example, here again some of the obtained effect sizes along with their 95% confidence intervals from these 20 repetitions fail to capture the real effect of treatment (i.e., .5), as indicated in red. And our obtained results (i.e., $d$ = .87; 95% CI [.33, 1.40]) could be "red" results, as is the case in four of the 20 repetitions here. Again, while in the long-run, the Frequentist procedure is correct (i.e., contains the true effect) in 95% of infinitely many repetitions of such a study, this assurance does not mean that our single obtained CI from our single study on Type III conditionals contains the true effect of the treatment with 95% certainty. Now imagine what a formidable challenge as well as confusion this might create for a researcher wanting to interpret the single obtained effect size, and its 95% confidence interval; certainty in a long-run procedure rather than in the single obtained result. Certainly, here the only *critical* question of interest is: What is the *real* effect of the L2 treatment on improving high-intermediate EFL learners' explicit knowledge of Type III conditionals?

Once again, the Bayesian method begins by asking our researcher about her/his expectation regarding the range of effect sizes reported in the previous research or the general domain of L2 research. We take a general approach here which appeals to a broader domain of L2 research. This makes such a Bayesian approach broadly accessible and provides a default prior distribution for Cohen's *d* effect size applicable to a wide range of domains in L2 research. To do so, we first draw on the results of Author and Author (xxxx-a) who studied the magnitude of Cohen's *d* effect size in 346 primary L2 studies and 91 meta-analyses of L2. The researchers found that *d* values in L2 research could often be as large as +1. Even so, conservativism dictates that one takes a neutral position and consider that Cohen's *d* effect size theoretically can be positive and negative. As such, it is safer to consider that the expected sizes of effect could be as large as reported by Author and Author (xxxx-a) in either a positive or negative direction (i.e., –1 and +1). Now that the range of likely effect sizes are at hand, it is time to assign higher weights to our expected range and successively lower weights to other effect size values outside this range. We use a "*Cauchy*" (named so in Augustin-Louis Cauchy's honor) weighting scheme to achieve this. A Cauchy weighting scheme, to be shown shortly, puts higher weights on the values of effect size between –1 and +1 than does the more familiar standard normal weighting scheme (see Rouder, Speckman, Sun, Morey, & Iverson, 2009). The technical specifics of the resultant prior distribution of effect sizes following the Cauchy weighting scheme are well documented (Ly, Verhagen, & Wagenmakers, 2016; Rouder, Morey, Verhagen, Province, & Wagenmakers, 2016). It is, however, worth noting that our prior distribution of effect sizes has a width (akin to standard deviation) of ".707", and is centered at "0" (i.e., our neutral position between positive and negative effect size values). Figure 17 shows this recommended prior distribution of effect sizes in which the area between –1 and +1 has received the highest weights (to explore Figure 17

see https://github.com/izeh/i/blob/master/d.r). Note that in theory, Cohen's *d* effect size has no bound. That is, it can be infinitely large in either direction. However, we can all agree that effect sizes beyond ± 6 are very unlikely. Thus, the largest values of effect size displayed in Figure 17 are ± 6 with two ± ∞ signs indicating the theoretical bounds of Cohen's *d* effect size.

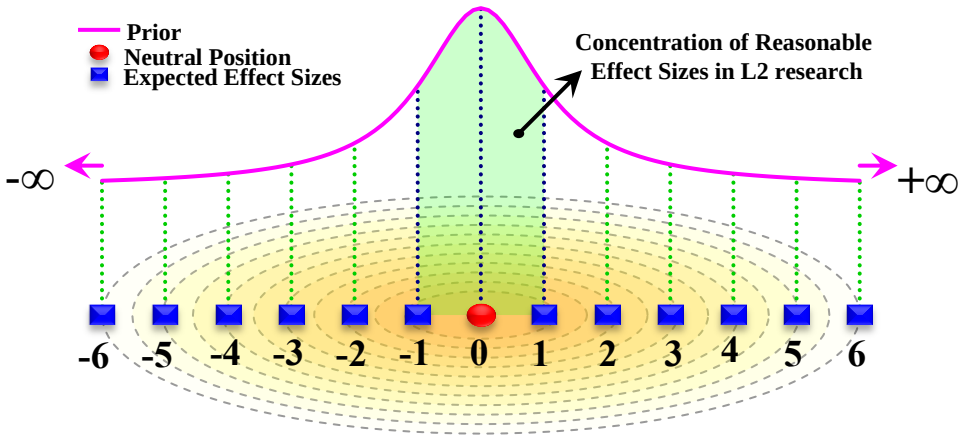


*Figure 17.* Recommended prior distribution for Cohen's *d* effect size in L2 research informed by Author and Author (xxxx-a).

In Figure 17, the dashed oval lines represent the main weighting domain of the prior. That is, the domain within which the possible effect size values in L2 research (e.g., −6 to +6) could receive various amounts of weight. The yellow color that spreads out from within the center of the dashed oval lines fades away as we move toward the large values of effect in the tails. This is to emphasize the fact that as we move from our neutral position (i.e., "0") toward the tails, the weights assigned to the individual effect size values successively decrease.

With the prior specified, the next steps involve determining the likelihood, applying the Bayesian mantra (Equation 1) to arrive at the posterior (i.e., the Bayesian result), and then obtaining a credible interval for the effect size to help the final interpretation. However, given the wide application of t-tests in real L2 research and the challenges inherent in learning to use new

statistical applications that permit Bayesian analyses, here we introduce a free, point-and-click, web-accessed software package developed by the first author of the present study to automate these processes. This software package is found at https://izeh.shinyapps.io/iiii/. The software will painlessly provide the posterior and the credible interval for effect sizes for the three, common t-test designs described above. For wider flexibility in terms of using a variety of different priors and robustness checks, we also provide easy to use R functions. The software has additional Bayesian capabilities that enable performing Bayesian hypothesis testing, and replacing *p*-values with a Bayesian alternative known as a *Bayes Factor*. The issue of Bayes Factors/Bayesian Hypothesis Testing/Model Selection, however, falls outside the scope of the present study (for details see Author et al, xxxx). Figure 18 provides a snapshot of the main panel of the software.

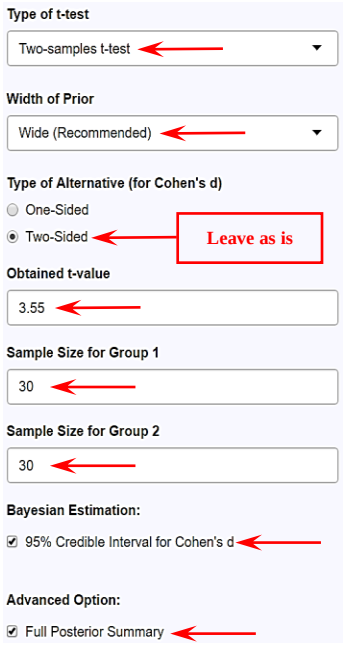


*Figure 18.* A snapshot of the "Bayesian for t-tests" software. The red arrows indicate the settings used for the example in the text.

To use the software in our example, we do not need to provide the raw data shown in Figure 7. Rather, only the following information is required: (1) the type of t-test, (2) the width of the prior, (3) the obtained t-value, (4) the groups' sample sizes. These four pieces of information for our example study on Type III conditionals are indicated by *red arrows* in Figure 18. Figure 19 (explore the software output) summarizes the software's Bayesian result (i.e., posterior) for our running example.

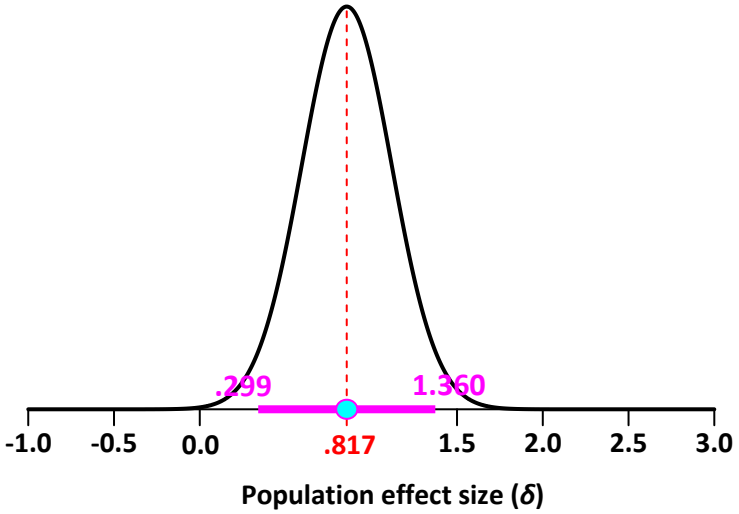


*Figure 19.* Posterior results for the effect of an L2 treatment on improving 60 high-intermediate EFL learners' explicit knowledge of Type III conditionals (see the text).

Now our 95% Bayesian credible interval can help us think about the *real* effect of our treatment, as measured in terms of Cohen's *d* effect size, on improving high-intermediate EFL learners' explicit knowledge of Type III conditionals. Here, we can directly state that there is 95% probability that the *real* effect size for our treatment could range from .299 to 1.360. One of the appealing features of the software is that it automatically provides the corresponding Frequentist results along with the Bayesian results. For our example, the Frequentist 95% confidence interval limits for effect size are: [.380, 1.445]. Again, this confidence interval is theoretically only one of the infinitely many possible confidence intervals that can result from repetitions of our study,

and thus we cannot take its 95% confidence level as 95% certainty that this single obtained confidence interval contains the true effect of our treatment on improving explicit knowledge of Type III conditionals. Research shows that the temptation to erroneously interpret a Frequentist confidence interval as if it is a Bayesian credible interval is considerably high despite the fact that such an interpretation is not permissible under the Frequentist framework (Albert, 2009; Gelman et al., 2014; Kruschke, 2015; McElreath, 2016).

**Putting priors on Cohen's *d* effect size to the test**

As noted earlier, it is always recommended and useful to test the robustness of the Bayesian result (i.e., posterior) for any research parameter against the choice of prior, and effect size is no exception. Here again the nature of effect size and type of priors commonly used with it should govern how one might want to go about choosing priors for such sensitivity analyses. Specially, the intrinsic meaning of effect size as a research result should guide us in determining (a) how wide priors on an effect size could be, and (b) where to center the priors as a pivot point. Given these two considerations, one possible way to start the robustness analysis is to use different families of priors that cover a realistic range for effect size (e.g., −6 to +6) while they might differ in distributing their weight over this realistic range. Note that too wide or too narrow specifications of prior in the case of effect size could easily lead to the assignment of undue weights to values for effect size that might not realistically need such amounts of weight. For example, prior specifications for effect size that are too narrow may unrealistically ignore effect sizes that are slightly larger than |1|, and too wide of a specification may give fairly large effect size (e.g., > |3|) more weight than required. Let us use two families of priors, namely Normal, and Cauchy. These two prior families for effect size (Cohen's *d*) could be used when they are each pivoted at "0" (a neutral position) and their width set to "1" and "1.25" (two reasonably

wider settings compared to .707 used in the previous section). This plan leads to four different prior specifications: *Cauchy*(0, 1), *Normal*(0, 1), *Cauchy*(0, 1.25), *Normal*(0, 1.25).

As in the case of proportions in the previous section, the goal is to evaluate the robustness of the Bayesian result obtained in Figure 19 under four different prior specifications. To do this, we can use the function "`d.bayes`" which uses the t-value (`t`), group samples sizes (`n1` or/and `n2`), pivot point for priors (`m`), and the width of prior (`s`):

```
d.bayes(t = 3.55, n1 = 30, n2 = 30, m = 0, s = rep(c(1, 1.25), 2),
dist.name = rep(c("dcauchy","dnorm"), each = 2), scale = .6, top = .9)
```

The result of our analyses is illustrated in Figure 20.

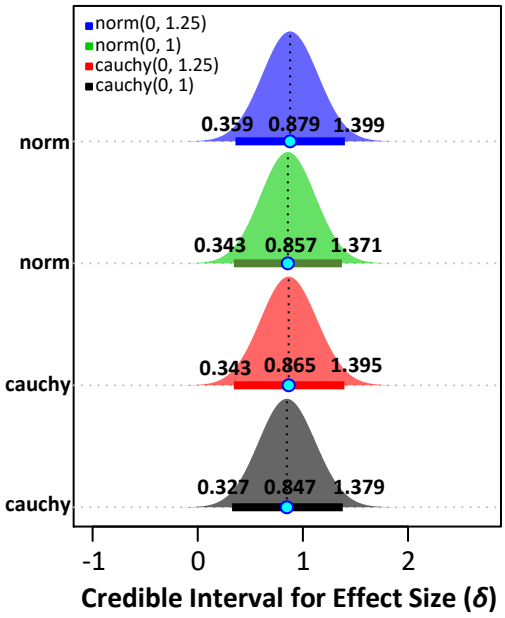


*Figure 20*. The credible intervals under different families and specifications of prior

As can be seen, the 95% credible intervals under these different priors still range from ~.3 to ~1.4. Thus, it is safe to believe that under such reasonably different prior specifications (i.e., wider and of different families), our Bayesian result for our study on type III conditional is reasonably stable. The interested reader may use other families of priors such as standard *t*

distribution (i.e., `dist.name = "dt", s = 0`) with a few degrees of freedom (e.g., `m = 5`) to see that the credible intervals are still robust to this other reasonable expression of prior knowledge on the effect size in the example of Type III conditionals.

**Doing Bayesian Estimation Using Published Findings: Two actual examples**

The discussion in the previous section should imply the ease with which full Bayesian estimation of effect sizes can be performed even on previously published studies. As an even more concrete example, consider Gurzynski-Weiss and Baralt (2014). Using a pre-post design, one of the questions that the authors investigated was the effect of the interaction mode (i.e., computer-mediated communication [CMC] vs. face-to-face [FTF]) when providing 24 intermediate-level learners of Spanish as a foreign language (SFL) with opportunities to modify their output during interactional feedback episodes with their teacher. After eliciting their data via stimulated recall protocols (see Mackey & Gass, 2016), the authors conducted a paired-samples t-test to answer their research question, finding $t$ (23) = 5.03, with descriptive results favoring the FTF environment. This is enough information for us to perform a secondary Bayesian estimation of the effect size on this study using the default prior proposed in the previous section. Changing the software settings to a paired-samples t-test, and inputting the sample size of 24, and the obtained t-value of 5.03 will provide us with the result in Figure 21 (explore the software output).

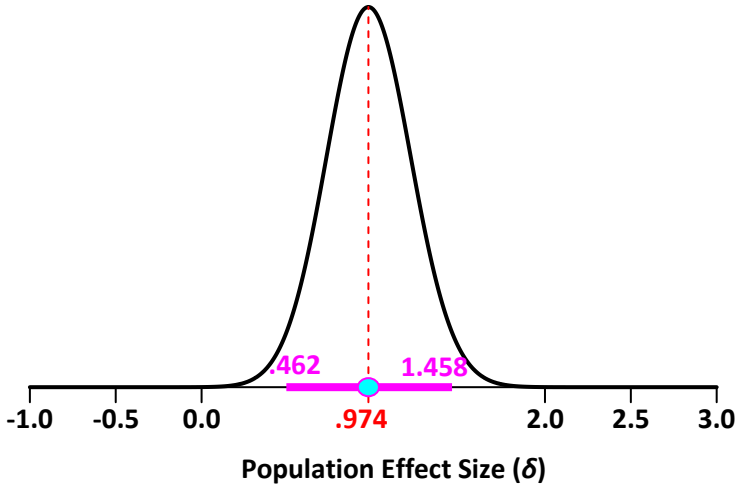


*Figure 21*. Posterior distribution for the effect size found in Gurzynski-Weiss and Baralt (2014) for the superiority of FTF environment over CMC environment in affording more opportunities for modified output (see the text).

Succinctly put, if Gurzynski-Weiss and Baralt (2014) had conducted a Bayesian estimation for their study, they could have interpreted their results as directly and concisely as follows: there is 95% probability that the *real* superiority of the FTF over CMC in providing more opportunities for intermediate SFL learners to modify their output is quantified by Cohen's *d* estimates ranging between .462 and 1.458. Although not reported in Gurzynski-Weiss and Baralt (2014), using the software, the corresponding 95% confidence interval for their effect size, which is subject to a Frequentist interpretation, is: [.522, 1.516]. We encourage the informed reader to perform various robustness analyses on these results following our demonstration in the previous section.

Performing a secondary Bayesian analysis on one or more previously published studies is not only advantageous in providing a Bayesian interpretation of the previous research findings but also in effectively informing (as a cumulative prior) a future replication study. Recall from our previous discussions that *yesterday's posterior is today's prior* (see Lindley, 2000). For example, suppose previous research has shown that the advantage of FTF environments over CMC environments has been found to be fluctuating in three previous studies. More specifically,

in the first study with $n$ = 44, the result has indicated a smaller advantage for FTF over CMC ($t$ (43) = 2.36), for the second replication study with $n$ = 36 the result shows a moderate advantage ($t$ (35) = 3.39), and the third replication study with $n$ = 52 found a small advantage for FTF over CMC ($t$ (51) = 1.59). We can use these studies' results together as prior for Gurzynski-Weiss and Baralt (2014). To do so, we can use as knowledge base a *Cauchy*(0, 1) as a reasonably informative prior for effect size using the R function "`d.update`" from our repository:

```
d.update(t = c(2.36, 3.39, 1.59, 5.03), n1 = c(44, 36, 52, 24), scale = .21, top =
1.7, m = 0, s = 1, dist.name = "dcauchy", prior.scale = 2, margin = 1.5)
```

The result of this updating is shown in Figure 22.

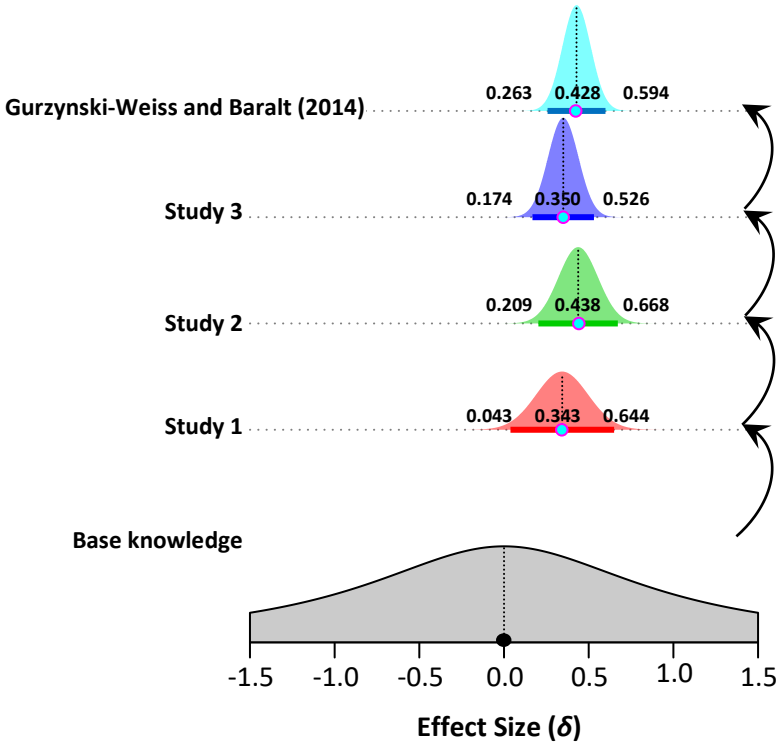


*Figure 22*. Step-wise Bayesian updating of three replication attempts to use them as prior for Gurzynski-Weiss and Baralt (2014)

When a researcher in a given subdomain of L2 research intends to adopt a Bayesian approach for her/his replication study, s/he can (a) perform secondary Bayesian analyses on

previous studies, regardless of whether the initial studies conducted a Bayesian estimation; (b) obtain the full posterior from those previous research works; and then (c) use the final posterior obtained in that step-wise updating process as the prior for her/his intended replication study. Such a practice is very consistent with the spirit of Bayesian methods which heavily rely on past research to inform a current replication study (see Note).

**Bayesian tests for correlational designs**

In addition to t-test designs, applied linguistics often employ correlational designs to examine possible relationships between various variables of interest (Author & Author, XXXX). In such cases, the focus is rightly on estimating the size of this relationship or effect. It is worth noting that linear bivariate correlations (e.g., Pearson *r*) are, themselves, a type of effect size (Author & Author, XXXX; Grissom & Kim, 2012; Thompson, 2006). Therefore, a Bayesian approach to estimating correlations aims to, as always, arrive at a Bayesian result (i.e., posterior) for this effect size.

Let us again use an actual research example through which the use of Bayesian correlations can be demonstrated. Ahmadian (2012) used a correlational design with 36 intermediate EFL (English as a Foreign Language) learners to examine the relationship between Working Memory Capacity (WMC), and measures of oral language performance (i.e., CAF; complexity, accuracy, and fluency) under Task-Based Careful Online Planning (COLP) condition (i.e., ample time for task completion with no initial time allowed for pre-task planning). Here we focus on the relationship between participants' WMC and production of fluent language in an oral narrative task under the COLP condition. Ahmadian used two sub-measures to measure fluency, namely number of syllables produced per minute (SPM), and number of meaningful syllables per minute (MSPM). He then correlated WMC measured via a

working memory span task (see Ahmadian, 2012, p. 168) with the two sub-measures of SPM ($r$ = .463) and MSPM ($r$ = .321). While Ahmadian found these two relationships to be statistically significant (i.e., $p < .05$), he did not provide any estimates of the generalizability of his results. Thus, we may wonder about the actual magnitude of these two relationships in the population of intermediate EFL learners? The R function "`cor.bayes`" can provide a Bayesian answer to this question. The function, by default, uses a type of scaled normal prior distribution that is bound between –1 and +1, centered at "0" (to indicate neutrality), and that has a standard deviation (i.e., width) of ".707". This prior specification is almost non-informative, and can be visually inspected using the following R command:

```
cor.bayes(prior.mean = 0, prior.sd = .707, show.prior = TRUE)
```

This prior distribution is also graphically shown in Figure 23.

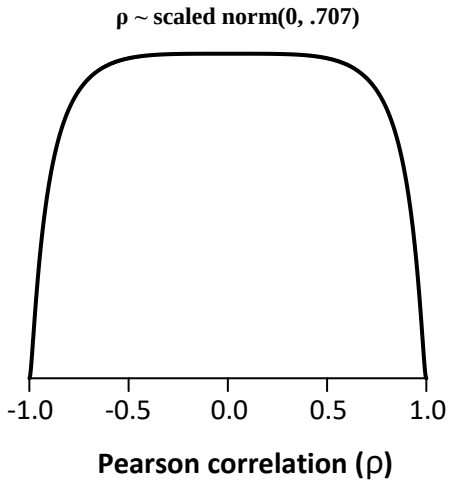


*Figure 23*. A prior distribution for Pearson correlation

We can insert the two correlations (`r`) that Ahmadian found in his study along with the number of participants (`n`) into the R function "`cor.bayes`" to obtain Bayesian estimates of the two relationships in question:

```
cor.bayes(r = c(.463, .321), n = 36, top = 1.4, scale = .3)
```

Figure 24 displays the visual result of these Bayesian estimations.

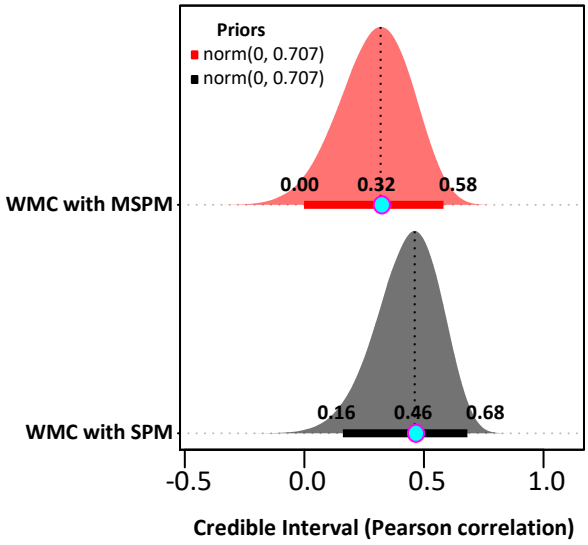


*Figure 24*. Posterior distributions for two correlations in Ahmadian (2012)

Inspecting our 95% Bayesian high-density credible intervals for Ahmadian's study, there is 95% probability that the magnitude of the relationship between WMC and SPM in the population of intermediate EFL learners can be as small as .16 and as large as .68. Also, the relationship between WMC and MSPM in the population of EFL learners could be estimated to be as small as nearly 0 to as large as .58. Also, note that as before, the Bayesian estimates for Ahmadian's study were obtained without needing to access the raw data used in the original study. Thus, again, Bayesian correlations can be easily carried out on previously published literature. Additionally, both the prior mean (`prior.mean`), and prior standard deviation (`prior.sd`) may be adjusted to match any defensible expectation for the magnitude of relationship between any two variables by the researcher.

**Modern Bayesian Machinery: Moving beyond proportions, t-tests, and correlations**

We have sought throughout this paper to introduce Bayesian methods in a manner that will be as clear and accessible to as many L2 researchers as possible. Our examples have intentionally involved simple scenarios. However, Bayesian methods are also very well suited to more sophisticated designs and analyses. In this final section, we introduce very briefly some of the potential of Bayesian analyses in such instances.

### *Markov Chain Monte Carlo (MCMC) and Bayes*

Modern Bayesian statistics is often characterized by its ability to approximate complex (e.g., multi-dimensional in shape) posteriors with many unknown parameters (e.g., multi-level models). In order to do so, many Bayesian analyses make use of a class of sampling techniques called *Markov Chain Monte Carlo* (MCMC). Detailed explanation of this sampling algorithm and its various offshoots which produce a *Markov Chain* falls outside the scope of the present paper (see Brooks, Gelman, Jones, & Meng, 2011; Gelman et al., 2014). We would simply note, though, that the introduction of MCMC sampling algorithms is often credited as giving birth to modern Bayesian methods, which are generally computationally intensive. We would also note that, while highly powerful and useful for approximating multi-parameter posteriors, a Markov chain must always be checked for its accuracy (i.e., success in approximating the posterior distribution). This being the case, several modern Bayesian software packages, by default, run multiple randomly initialized Markov chains in parallel for this precise purpose (see Depaoli & van de Schoot, 2017).

### *Bayesian Regression*

As discussed in Author and Author (XXXX), it is common for researchers to build models that explain or predict a phenomenon of interest using linear regression. As we might expect, linear regression, like the techniques discussed thus far, is quite possible and informative

using Bayesian methods. Given the space limitations, we only briefly discuss the case of Bayesian simple linear regression. More complete coverage of Bayesian regression is provided in many sources on Bayesian statistics (e.g., Gelman et al., 2014; Kruschke, 2015; McElreath, 2016).

Suppose we want to know if language proficiency measured via TOEFL iBT scores of 60 advanced EFL learners can be predicted by their Language Analytic Ability (LAA) as measured on the *Words in Sentences* subset of the Modern Language Aptitude Test (MLAT). Having collected the data, we might then use linear regression to find a linear model to relate LAA to language proficiency. In the current example, we know that TOEFL iBT scores from a large group of test takers approximately form a normal, bell-shaped curve (Wang, Eignor, & Enright, 2008). Therefore, it is reasonable to assume that our EFL learners' TOEFL iBT scores could similarly belong (symbolically denoted by "~") to a large, normally-distributed population. Essentially, this line of reasoning indicates that the likelihood function of our proficiency data could be a normal one, with some mean (center) and some standard deviation (width) determining its exact shape:

$$prof_i \sim Normal(mean_i, sd) \qquad Likelihood \qquad (2)$$

The job of the linear regression model is to indicate that the average proficiency (i.e., $mean_i$) of EFL learners' normal likelihood can linearly depend on learners' Language Analytic Ability (LAA) score:

$$mean_i = \alpha + (\beta \times LAA_i) \qquad Linear\ Model \qquad (3)$$

The linearity of this relationship is indicated by our linear model's use of only an additive constant called the intercept (α) and a multiplicative constant called the slope (β) to relate LAA scores to the average proficiency scores. For the sake of clarity, one can take (3) and substitute it

for the average proficiency ($mean_i$) in (2) to make it obvious that the likelihood of the proficiency data is now dependent on LAA scores:

$$prof_i \sim Normal(\alpha + (\beta \times LAA_i), sd) \quad (4)$$

So far, nothing Bayesian has occurred. The role of Bayesian inference starts once we set out to find the unknown parameters in our overall analysis. Based on the likelihood of our dependent variable (proficiency) (2), and our linear model (3), we have three unknown parameters. First, we do not know how wide (i.e., *sd* also denoted by the Greek letter "σ"; sigma) our normal distribution of proficiency scores for the EFL learners in the likelihood is. Second, we do not know the value of the intercept (α). And third, the value of the slope (β) is unknown to us as well. Instead of thinking that there is one true answer to each of these three unknown parameters, we can follow the Bayesian approach by assigning to each a prior. As suggested by Stan Development Team (2018), a weakly informative choice for the intercept (α) may be a wide normal distribution centered at "0", and standard deviation of "10". For the slope (β), the prior is suggested to be centered at "0", but narrower (e.g., width of 2.5) in width. For *sd*, one suggestion is to use a wide (e.g., width of 100) Cauchy prior that starts from 0 (*sd* cannot be less than "0"). The dataset "`prof`" which comes with our suite of R functions carries simulated TOEFL iBT scores of 60 advanced EFL learners along with their LAA scores suitable for demonstrating a Bayesian linear regression. We suggest running Bayesian linear regression models using the R function `stan_glm` from the `rstanarm` package (Stan Development Team, 2018) which has already been automatically installed on the reader's computer (see the Online Supplementary Document for a list of other software packages):

```
fit <- stan_glm(TOEFL ~ LAA, data = prof,                # dataset "prof"
       prior_intercept = normal(0, 10, autoscale = F),   # prior on intercept (α)
```

```
    prior = normal(0, 2.5, autoscale = F),   # prior on slope (β)
prior_aux = cauchy(0, 100, autoscale = F))   # prior on sd (σ)
```

As shown in Table 2, the summary of the results is provided using `summary(fit, digits = 2)`. These summary results are descriptions of our three unknown parameters' posterior distributions along with their 95% credible intervals.

Table 2. Posterior summary for the Bayesian linear regression

| **Estimates** | **Mean** | **Lower** | **Upper** | **Diagnostics (Rhat*)** |
|---|---|---|---|---|
| Intercept (α) | 88.84 | 82.04 | 95.66 | 1.00 |
| LAA (β) | .47 | .25 | .68 | 1.00 |
| Sigma (*sd*) | 3.65 | 3.06 | 4. 43 | 1.00 |

*Note.* *The **Rhat** statistic (Gelman et al., 2014, p. 285) is one of the several statistics to check if the Markov chains used to generate the posteriors of our three unknown parameters have reached an equilibrium state. An Rhat statistic below 1.1 is generally considered acceptable (also see Depaoli & van de Schoot, 2017).

The interpretation of the regression analysis is straightforward. Let us start with the β's posterior mean of .47. In this case, for every 1-point increase in an advanced EFL learner's LAA score, s/he is expected to perform .47 points higher on the proficiency measure. The posterior mean for α (the intercept) of 88.84 expresses the expected proficiency score for EFL learners whose LAA score is 0. As applied linguists, we might find it odd to think that there might be advanced EFL learners who might have absolutely no Language Analytic Ability. As a result, the value of the intercept in its raw form (i.e., without centering the predictor) is often ignored for interpretive purposes (see Author & Author, XXXX). The existence of *Lower* and *Upper* values in Table 2 reminds us that the intercept (α) of 88.84 and the slope (β) of .47 used to draw the

regression line are only one set of likely values for predicting EFL learners' TOEFL iBT scores from their LAA scores. When we consider other likely values for the intercept (α) and slope (β) from their posteriors, many other possible regression lines begin to appear to form a halo around the original regression line. The R function "`predict.bayes`" from our suite of R functions makes this Bayesian notion more obvious:

```
predict.bayes(fit)
```

The result of our `predict.bayes` function is shown in Figure 25.

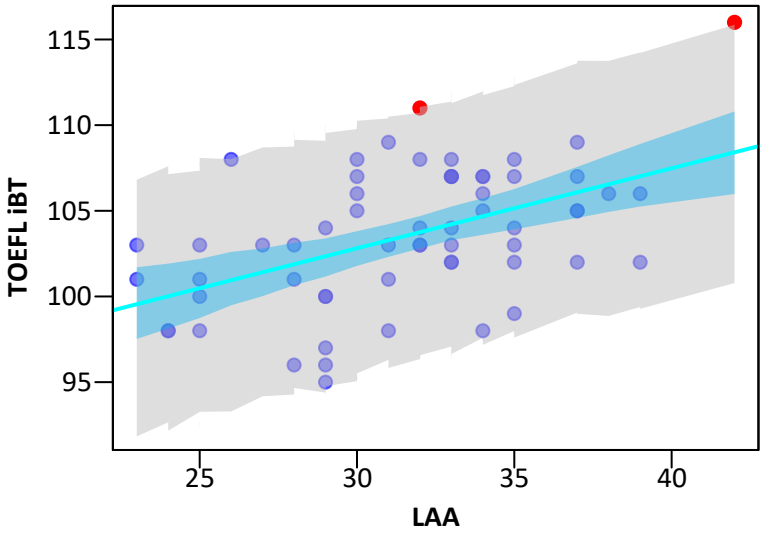


*Figure 25.* Regression line and prediction uncertainties in Bayesian regression

The light-blue area around the regression line visually represents the most likely positions of the regression line in the actual population of advanced EFL learners. In addition to the likely positions of the regression line, our Bayesian regression model can predict the most likely proficiency scores of advanced EFL learners based on their specific LAA scores. The grey-colored area offers such a prediction. For instance, if we are interested in predicting the most likely proficiency scores of advanced EFL learners who possess a medium level of Language Analytic Ability (i.e., an LAA score of about 23 on the *Words in Sentences* subset of MLAT),

provided that we have actual learners in our own collected data with an LAA score of 23, we can cut a vertical slice at the LAA score of 23 out of the grey area. Our R function `predict.case` is designed for this purpose:

```
predict.case(fit, 23)
```

Since the TOEFL iBT test used to measure proficiency only results in integer scores (e.g., 92 but not 92.03), one may like to view the predictive results in the rounded form to get more practically usable predictions. The result of our prediction is displayed in Figure 26. These results indicate that there is 95% probability that an advanced EFL learner with a medium level of Language Analytic Ability could score as low as ~92 and as high as ~107 on the proficiency measure (TOEFL iBT).

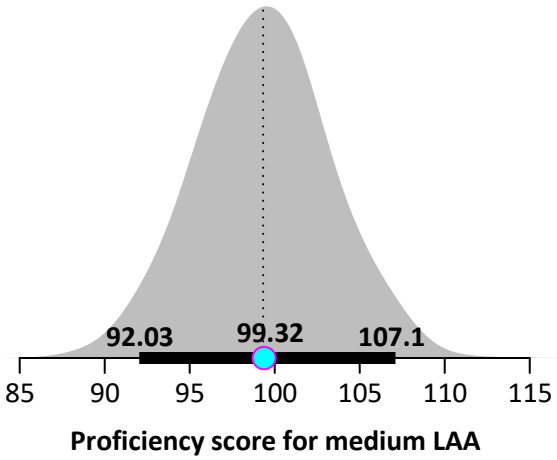


*Figure 26.* Regression line and prediction uncertainties in Bayesian regression

Finally, we can quickly check the overall fit of the model to the data using the Bayesian version of R-squared (i.e., $R^2$). Using our R function "`R2.bayes`": `R2.bayes(fit, top = 1.2)` we realize that there is 95% probability that a model such as the one we just fit could produce an R-squared of as small as ~8.18% and as large as ~39.55% in the actual population of advanced EFL learners. These interval estimates describe the predictive power of Language

Analytic Ability in relation to proficiency and are almost always absent in the Frequentist version of linear regression studies.

Given that no model works best for all participants in a study, one might want to know how well a model fits each participant recruited for the study. Our R function `case.fit.plot` provides a visual answer to this question:

```
case.fit.plot(fit)
```

Figure 27 presents the result of our person fit analysis.

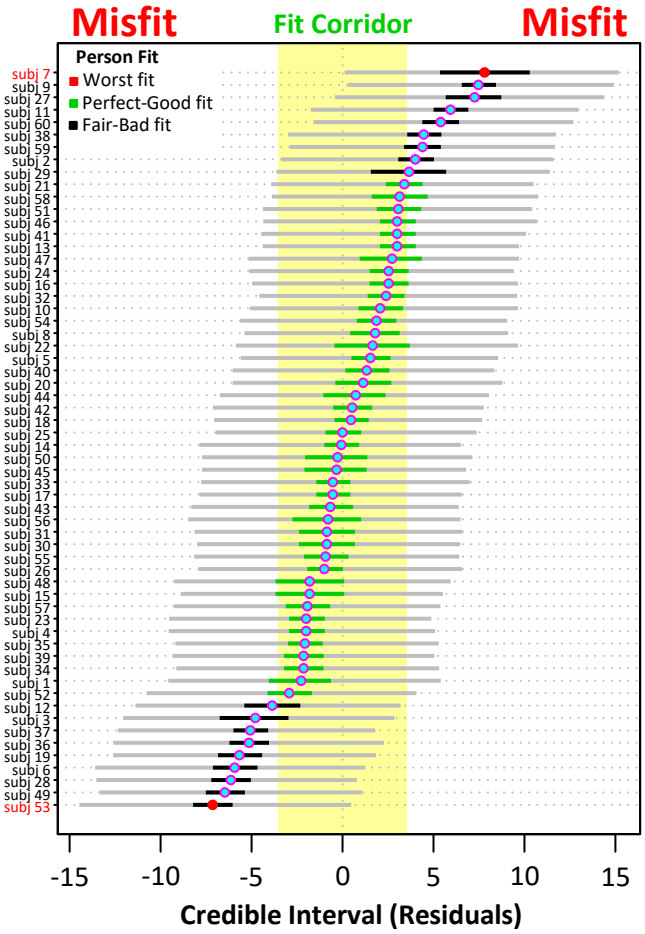


*Figure 27.* Person fit analysis for 60 advanced EFL learners after a Bayesian regression

As shown in Figure 27, the model has worked relatively well for several participants who have fallen within the yellow corridor (i.e., fit corridor). However, for participants outside the yellow

corridor, the model has worked less well, particularly for participants such as subj # 7 and subj # 53 who have been indicated in red.

**Conclusion**

There is a statistical view of the world that has long permeated the scientific literature. By the basic rules of this world, there are good reasons to believe what we report as "findings" from our studies might not represent the reality we are attempting to capture. To learn about that reality, however, two solutions exist.

The first solution relies on a procedure that assumes repeating one's exact same study ad infinitum, providing a specified certainty (e.g., 95%) in capturing the true effect in question from this long-run procedure (Frequentism). Under this approach, the interpretation of a single observed interval estimate (i.e., confidence interval) must be made in the context of the Frequentist procedure i.e., over long-run frequencies, 95% of the confidence intervals theoretically constructed in the process (see Figures 2 and 16) would contain the true population value and not in terms of the single interval estimate obtained (see Depaoli & van de Schoot, 2017; Thompson, 2006). This Frequentist interpretation likely escapes the awareness of many applied researchers.

The second solution, which we advocated in the present paper, translates the theoretical repetitions assumed in the Frequentist paradigm into a prior distribution. That is, a prior is a practical way for expressing defensible expectations for reality rather than thinking about reality in Frequentist terms.

By nature, reasonableness and conservatism must always govern the use of Bayesian statistics. Choices of priors must be transparent as they are an orderly form of knowledge presentation (Edwards, Lindman, & Savage, 1963). Decisions made at every step of the analyses

must be defensible. And researchers must routinely evaluate the robustness of the obtained results and report them to their audience (for a complete checklist of points to consider when conducting a Bayesian analysis see Depaoli & van de Schoot, 2017). However, we argue that with Bayesian methods taking a central stage in L2 research, we will enter a new era marked by (a) constructive criticisms and academic debates over key issues in the assessment and development of L2 theory, (b) more precise attention to past research findings to come up with defensible priors, and (c) a focus on meaningful research parameters worthy of being estimated (e.g., effect sizes).

These three advantages from Bayesian methods, we believe, best characterize the need for a "Bayesian revolution" in L2 research. Thus, we hope the applied, and non-technical approach that we adopted in this paper could be a first step for the field in that direction.

## Note

The broader framework for synthesizing outcomes of multiple studies when differences between studies (due to differences among sampled participants in the studies and differences in treatments, settings etc.) also exist is the form of random-effects meta-analysis (Cooper, Hedges, & Valentine, 2009). Bayesian methods are capable of seamlessly handling random-effects meta-analysis even in the face of a limited number of primary studies available, a problem often restricting the use of random-effects meta-analysis under the Frequentist framework. The topic of Bayesian meta-analysis falls outside the scope of the present treatment. The interested reader is referred to Berry, Carlin, Lee, and Müller (2011, Sec. 2.4), Smith, Spiegelhalter, and Thomas (1995), Spiegelhalter, Abrams, and Myles (2004, Ch. 8), Stangl and Berry (2000), and Sutton and Abrams (2001) for a foundational introduction. The R packages "`bayesmeta`" (Röver, 2017)

and "`bmeta`" (Ding & Baio, 2016) both provide efficient implementation of Bayesian random-effects meta-analyses for a variety of study outcome metrics (e.g., standardized mean difference effect size).

**References**

Ahmadian, M. J. (2012). The relationship between working memory capacity and L2 oral performance under task-based careful online planning condition. *TESOL Quarterly, 46*(1), 165-175.

Albert, J. (2009). *Bayesian Computation With R* (2nd ed.). New York: Springer.

Author. (xxxx).

Author, & Author. (xxxx-a).

Author & Author. (xxxx-b).

Author et al. (xxxx).

Bedore, L. M., Peña, E. D., Joyner, D., & Macken, C. (2011). Parent and teacher rating of bilingual language proficiency and language development concerns. *International Journal of Bilingual Education and Bilingualism, 14*(5), 489-511.

Berry, S. M., Carlin, B. P., Lee, J. J., & Müller, P. (2011). *Bayesian adaptive methods for clinical trials*. NY, New York: CRC press.

Brooks, S., Gelman, A., Jones, G., & Meng, X. L. (2011). *Handbook of markov chain monte carlo*. NY, New York: CRC press.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Boston: Houghton Mifflin.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*(12), 1304 -1312.

Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis*. New York: Russell Sage Foundation.

DeKeyser, R. (2015). Skill acquisition theory. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 94-112). London: Routledge.

Depaoli, S., & van de Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: The WAMBS-Checklist. *Psychological Methods, 22*(2), 240.

Dienes, Z., & Mclatchie, N. (2017). Four reasons to prefer Bayesian analyses over significance testing. *Psychonomic Bulletin & Review*, 1-12. doi:10.3758/s13423-017-1266-z

Ding, T., & Baio, G. (2016). bmeta: Bayesian Meta-Analysis and Meta-Regression. R package version 0.1.2. available at: https://CRAN.R-project.org/package=bmeta.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70*(3), 193-242.

Ellis, R. (2009). Measuring implicit and explicit knowledge of a second language. In R. Ellis, S. Loewen, C. Elder, J. Philp, & H. Reinders (Eds.), *Implicit and explicit knowledge in second language learning, testing and teaching* (pp. 31–64). Bristol: Multilingual Matters.

Etz, A., & Vandekerckhove, J. (in press). Introduction to Bayesian Inference for Psychology. *Psychonomic Bulletin & Review*.

Farruggio, P. (2010). Latino immigrant parents' views of bilingual education as a vehicle for heritage preservation. *Journal of Latinos and Education, 9*(1), 3-21.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian data analysis* (2nd Ed.): Chapman & Hall/CRC Boca Raton, FL, USA.

Gurzynski-Weiss, L., & Baralt, M. (2014). Exploring learner perception and use of task-based interactional feedback in FTF and CMC modes. *Studies in Second Language Acquisition, 36*(1), 1-37.

Izumi, S., Bigelow, M., Fujiwara, M., & Fearnow, S. (1999). Testing the output hypothesis: Effects of output on noticing and second language acquisition. *Studies in Second Language Acquisition, 21*(3), 421-452.

Jaynes, E. T. (2003). *Probability theory: The logic of science*: Cambridge university press.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). New York: Oxford University Press.

Kruschke, J. K. (2015). *Doing Bayesian data analysis* (2nd ed.). Boston: Academic Press.

Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The time has come: Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods, 15*(4), 722-752.

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 and Review*, 1-29. doi:10.3758/s13423-016-1221-4

Larson-Hall, J. (2016). *A guide to doing statistics in second language research using SPSS and R* (2nd ed.). New York: Routledge.

Levine, T. R., & Hullett, C. R. (2002). Eta squared, partial eta squared, and misreporting of effect size in communication research. *Human Communication Research, 28*(4), 612-625.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association, 103*(481), 410-423.

Linck, J. A., & Cunnings, I. (2015). The utility and application of mixed-effects models in second language research. *Language Learning, 65*(S1), 185-207. doi:10.1111/lang.12117

Lindley, D. V. (2000). The philosophy of statistics. *The Statistician, 49*(3), 293-337.

Ly, A., Verhagen, J., & Wagenmakers, E.-J. (2016). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology, 72*, 19-32. doi:10.1016/j.jmp.2015.06.004

Lyster, R., & Sato, M. (2013). Skill acquisition theory and the role of practice in L2 development. In P. García Mayo, M. Gutierrez-Mangado, & M. Martínez Adrián (Eds.), *Contemporary approaches to second language acquisition* (pp. 71-92). Amsterdam: Benjamins.

Mackey, A., & Gass, S. M. (2016). *Second language research: Methodology and design* (2nd ed.). New York: Routledge.

Marsden, E. J., Morgan-Short, K., Thompson, S., & Abugaber, D. (in press). Replication in second language research: Narrative and systematic reviews, and recommendations for the field. *Language Learning*.

McElreath, R. (2016). *Statistical rethinking: A Bayesian course with examples in R and Stan.* NY, New York: CRC Press.

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology, 72*, 6-18.

Pierce, C. A., Block, R. A., & Aguinis, H. (2004). Cautionary Note on Reporting Eta-Squared Values from Multifactor ANOVA Designs. *Educational and Psychological Measurement, 64*(6), 916-924. doi:10.1177/0013164404264848

Porte, G. (2012). *Replication research in applied linguistics*: Cambridge University Press.

Ramos, F. (2007). What do parents think of two-way bilingual education? An analysis of responses. *Journal of Latinos and Education, 6*(2), 139-150.

Rouder, J., Morey, R., Verhagen, J., Province, J., & Wagenmakers, E. J. (2016). Is there a free lunch in inference? *Topics in Cognitive Science, 8*(3), 520-547. doi:10.1111/tops.12214

Rouder, J. N., Speckman, P., Sun, D., Morey, R., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin and Review*, *16*(2), 225-237.

Röver, C. (2017). Bayesian random-effects meta-analysis using the bayesmeta R package. *arXiv preprint arXiv:1711.08683*.

Salkind, N. J. (2010). *Encyclopedia of research design*. CA, Thousand Oaks: Sage.

Shintani, N., Ellis, R., & Suzuki, W. (2014). Effects of written feedback and revision on learners' accuracy in using two English grammatical structures. *Language Learning*, *64*(1), 103-131.

Smith, T. C., Spiegelhalter, D. J., & Thomas, A. (1995). Bayesian approaches to random-effects meta-analysis: a comparative study. *Statistics in Medicine, 14*(24), 2685-2699.

Spiegelhalter, D. J., Abrams, K. R., & Myles, J. P. (2004). *Bayesian approaches to clinical trials and health-care evaluation*. Sussex, England: Wiley.

Stan Development Team. (2018). Stan: A C++ library for probability and sampling. Available at: http://mc-stan.org.

Stangl, D., & Berry, D. A. (2000). *Meta-analysis in medicine and health policy*. New York: CRC Press.

Sutton, A. J., & Abrams, K. R. (2001). Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research, 10*(4), 277-303.

Thompson, B. (2006). *Foundations of behavioral statistics: An insight-based approach*. NY, New York: Guilford Press.

Wang, L., Eignor, D., & Enright, M. K. (2008). A final analysis. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.) *Building a validity argument for the Test of English as a Foreign Language*. New York: Routledge.