Bayesian Analysis as Alternative to Frequentist Methods: A Demonstration with Data from Language-Impaired Children's Relative Clause Processing

Yair Haendler, Romy Lassotta, Anne Adelt, Nicole Stadie, Frank Burchert, and Flavia Adani

1. Introduction

Experimental studies in linguistics, like other fields in the cognitive sciences, have traditionally used Frequentist analysis methods. In the Frequentist analysis, the researcher performs some statistical test (e.g., t-test, ANOVA, linear mixed-effects model) and is typically interested in the p-value associated with the test statistics of main effects and interactions. A significant effect, with a p-value smaller than 0.05, commonly brings the researcher to formulations like 'there is evidence that the two conditions under comparison are different'. If the p-value is equal to or bigger than 0.05, the typical conclusion is that there is 'no evidence for the difference between the two conditions'.

There are some misconceptions and misunderstandings concerning what the output of the statistical test in the Frequentist method means, and how inference should be drawn from it. Moreover, although the Frequentist method itself is not erroneous, some inherent characteristics of the method make it uninformative or even inappropriate in certain contexts. In recent years, a thorough discussion has taken place around how these shortcomings of the Frequentist method can be avoided using Bayesian analysis methods (Kruschke & Liddell, 2018; McElreath, 2016; Nicenboim et al., 2018). In this article, we explain and discuss these issues with specific reference to research in language development.

We will first define and explain basic concepts in the Frequentist approach. We will spell out the correct way of drawing inference when using this method and describe three main shortcomings. We will then explain what Bayesian analysis is and how it avoids these drawbacks. Next, we will present a study conducted with children with Developmental Language Disorder (DLD) and typically developing controls on the online processing of relative clauses in

* Yair Haendler, Université de Paris, yairhen@gmail.com; Romy Lassotta, pädquis Stiftung, r.lassotta@paedquis.de; Anne Adelt, Staatliche Berufsfachschule für Logopädie in Regensburg, anne.adelt@ukr.de; Nicole Stadie, University of Potsdam, nstadie@uni-potsdam.de; Frank Burchert, University of Potsdam, burchert@uni-potsdam.de; Flavia Adani, Freie Universität Berlin, flavia.adani@fu-berlin.de. We thank the children who participated in the study and their families. The study was supported by *Deutsche Forschungsgemeinschaft* grant AD 408/1-1.

 Proceedings of the 44th Boston University Conference on Language Development, ed. Megan M. Brown and Alexandra Kohut, 168-181. Somerville, MA: Cascadilla Press.

German, using these data to demonstrate the procedure of the proposed Bayesian analysis. A discussion about the two analysis methods will conclude the paper.

2. Frequentist analysis

To illustrate what happens in a Frequentist analysis, let us use an example. A researcher did a sentence-picture matching task to test children's comprehension of two sentence types, A and B. The hypothesis the researcher would like to (dis)confirm is that sentence A is easier than sentence B. The two relevant numbers are the group's mean accuracy on condition A (e.g., 0.78) and on condition B (e.g., 0.65). The two means are compared with a statistical test, say a t-test. The resulting p-value is 0.01 (t=2.56). The researcher writes her conclusion: "we found evidence that sentence A is easier than sentence B".

Concluding, based on a p-value, that there is or there is no evidence for a particular effect is widely common, but it is incorrect. Such conclusions ignore the fact that a p-value is conditional probability. The p-value gives us the probability of observing a test statistics at least as extreme as the one we observe in our experiment, conditional on the assumption that the Null Hypothesis is true (Vasishth & Nicenboim, 2016). The Null Hypothesis, made implicitly in the Frequentist Null Hypothesis Significance Testing (NHST) procedure, is the hypothesis that there is in reality no difference between sentence A and sentence B. The p-value thus gives us information about the Null Hypothesis (H_0), not about the researcher's hypothesis–the Alternative Hypothesis (H_1)–that the conditions do differ. This distinction is important. As researchers, we usually want to find information about $P(H_1 \mid \text{data})$, that is, the probability that our hypothesis, the Alternative, is correct given the data we collected. But the p-value gives us information about $P(\text{data} \mid H_0)$, which is the probability of observing the data given that the Null Hypothesis is true. Thus, a p-value only allows us to formulate conclusions concerning the Null Hypothesis, not about the Alternative. Our researcher should say "given the data, we found evidence that the Null Hypothesis is wrong". But that is still no evidence in favor of the Alternative.

To better illustrate why we cannot infer anything about the Alternative based on information related to the Null, we can use Vasishth & Nicenboim's (2016) example from Dienes (2011). When a shark bites one's head clean off, the probability of dying is 1, such that P(death | shark head bite) = 1. But shark bites are an extremely rare death cause. In fact, given a dead person, the probability that a shark has bitten their head off is close to zero, so P(shark head bite | death) = 0.000000001. Similarly, a p-value smaller than 0.05 tells us that the Null is wrong, not that the Alternative is true.

So far we have pointed out a problem that is related to how researchers use the Frequentist method, namely their mistaken way of drawing inference from its outcome. There are at least two additional shortcomings that are inherent to the Frequentist approach itself. The first is the fact that, with small amounts of data, we run the risk of incurring Type I and Type II errors (Vasishth & Nicenboim, 2016). The Frequentist approach is based on the idea of hypothetically repeating

the experiment over and over again. Under such repeated sampling, Type I Error ('false positive') is defined as the probability of incorrectly rejecting the Null Hypothesis whereas in reality it is true. Type II Error ('false negative') is the probability of failing to reject the Null whereas in reality it is false. With a small sample size, the Frequentist analysis results are likelier to simply reflect such errors, so inference based on them might be erroneous. With large amounts of data the situation is different, but this is almost never the case, since we rarely have hundreds of participants in our experiments, especially when working with special populations, like children with DLD.

The second problem is that complex Frequentist models often fail to converge (Kimball et al., 2019). This might occur partly due to too small amounts of data or also because the estimation of model parameters is not restricted enough (Schad et al., submitted). In the face of failed convergence, the researcher has little to do besides reducing model complexity by excluding random effects parameters. But a reduced model might be inappropriate in that it does not reflect the process by which the data were generated in the experiment (Sorensen et al., 2016). Moreover, a model with only random intercepts might lead to a higher Type I error rate (Barr et al., 2013; Matuschek et al., 2017).

In sum, the Frequentist approach has several shortcomings. One disadvantage is that researchers tend to draw unwarranted conclusions from its output (the same holds not only for p-values, but also for t-values, F-statistics, etc.; it is about the way of drawing inference, not about what kind of statistics is used; see Kruschke, 2015). To overcome this issue, we could in principle correct our inference-making and formulate conclusions concerning the Null, rather than the Alternative Hypothesis. But this means giving up on what we are interested in, namely inferring something about our hypothesis in light of the new data. Furthermore, the risk of Type I and Type II errors when having little data and convergence issues still remain. In the next section, we will describe what a Bayesian analysis is and how it avoids the discussed shortcomings of the Frequentist approach.

3. Bayesian analysis

Bayesian analysis is not a model or statistical test. It is a different way of looking at the process of data analysis and doing statistical inference. In fact, we could work with Bayesian equivalents of various tests: ANOVA, t-test, etc. Here we will use linear mixed-effects models. In terms of approaching the analysis procedure itself, there is an important conceptual difference between the Frequentist and Bayesian methods. Recall that in the former we get information about $P(\text{data} \mid H_0)$, that is, we infer something about the data under the assumption that the Null Hypothesis is true. By contrast, in a Bayesian analysis we get information about $P(H_1 \mid \text{data})$, enabling us to infer something about our hypothesis given the data at hand. Bayesian probability used to be called 'inverse probability', because we use the data to infer something on the theory, unlike the Frequentist approach in which we use theory to infer something about the data.

Hence, under the Bayesian approach we may formulate conclusions about evidence that there is or there is no difference between conditions. Thus, when we do statistical inference we may keep our (Bayesian) intuitions by using an appropriate inference approach. There is no Null Hypothesis in a Bayesian analysis, which is why inference is done directly on the experimenter's hypothesis or theory. Since the method is not based on hypothetical repeated sampling, there are no Type I or Type II errors, which makes this approach suitable even for small amounts of data. Finally, complex Bayesian models do not fail to converge with small amounts of data. Therefore, they allow more flexibility in terms of what model to use and how to structure it. These characteristics of the Bayesian approach make it a powerful alternative to the Frequentist approach. An additional advantage of the Bayesian analysis is the ability to integrate into the analysis knowledge from domain expertise and previously collected data.

There are three components in a Bayesian analysis: prior, likelihood and posterior. The *prior* expresses some belief or assumption about the outcome, which is formulated by the experimenter. The *likelihood* is the data we collected. The *posterior* is the outcome of the analysis. It is an evaluation, made by some compromise between the prior and the likelihood, that tells us how the initial belief or assumption (prior) should be updated given the data at hand (likelihood) (Kruschke, 2015). All three components should be thought of in terms of probability distributions, and as such they are expressed in the analysis.

In fact, Bayesian inference is probabilistic (Kruschke, 2015). It is about how likely it is to obtain a particular outcome (posterior) given some data and a prior. Since the outcome is not binary, but rather a probability distribution, the results might have various degrees of meaningfulness. This is another important contrast with the Frequentist approach, in which the outcome is binary: an effect or an interaction can be either significant ($p < 0.05$) or not ($p \geq 0.05$). It is important to emphasize that, under the Frequentist approach, there is no such thing as a "trend" or an effect that is "marginally/almost significant". Such formulations reflect once again our Bayesian intuitions that the results can have varying degrees of meaningfulness, but they are absolutely incompatible with Frequentist methods.

3.1. Priors

As mentioned, the prior reflects an assumption or belief concerning the data. But how can a researcher have an idea in advance about what the results will look like? Consider again the sentence-picture matching task on the comprehension of sentences A and B. Can the outcome be anything between $-\infty$ and ∞? Surely not. To start with, we know that the resulting mean difference between the two conditions will be on a response accuracy scale. So at the very least we can say that the range of possible outcomes will be anything between 0 and 1, where all potential outcomes are equally likely to occur (Figure 1a). This kind of prior is called *uninformative* or *flat prior*.

But in most cases we can restrict our assumption about the results. Even if we have no idea about which sentence might be easier, we could speculate that

the mean difference is 0.5 with some rather large standard deviation (Figure 1b). Such a prior, called *weakly informative prior*, reflects the fact that we have a vague idea about the outcome. An assumed mean of 0.5 reflects the vagueness of knowledge about which condition is easier; the large standard deviation reflects the uncertainty around this assumption. But this prior is still more restricted than the flat prior, in which any outcome is equally possible.

In many cases, we could restrict our assumptions even further and come up with a more precise distribution of the possible outcomes. Let us say that four other studies, some conducted in our own lab and some by colleagues, used a similar task to test similar sentences in a comparable population. We can use the results of these previous studies to formulate a prior distribution with a specific mean and standard deviation for the difference in accuracy between the two sentences (Figure 1c). This type of prior is called *informative prior*.

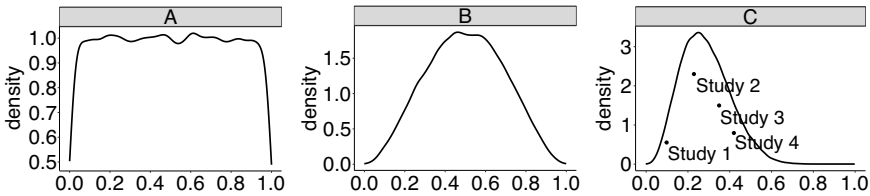

Figure 1. Possible prior distributions for the mean accuracy rate of the difference between two conditions: flat prior (A); weakly informative prior (B); informative prior based on four hypothetical previous studies (C).

Generally, flat priors are not recommended. On the practical level, they give a lot of probability mass to values that in reality are unlikely to occur. This might influence the results when we have little data, with the prior being dramatically determining for the posterior distribution (Nicenboim & Vasishth, 2016). Flat priors do not make much sense also on a conceptual level. We almost never start a new experiment without an idea about the potential outcome. On the contrary, in most cases our experiments are designed to answer a research question that emerged from previous experiments. Many studies use similar material and methods, making each result a useful component of the prior for the next experiment. But even without previous studies to rely on, for instance when we test this kind of material or use this method for the first time, we can still formulate a vague assumption about the outcome, opting for a weakly informative prior, which is preferable to a flat one.

4. The current study

The study we use to demonstrate the analysis procedure was designed to test the processing of German subject and object relative clauses with different types of embedded constituent (see Adelt et al., 2017 on adults with aphasia). Participants were children with a developmental language disorder (DLD) and typically developing controls. Previous research suggests that children with DLD

have more difficulties with object relatives than their typically developing peers (Contemori & Garraffa, 2010, a.o.). In some studies, children with DLD appear to have difficulties also with subject relatives (e.g., Stavrakaki, 2002). The present study aimed at testing children with DLD and typically developing controls on the processing of relative clauses disambiguated by case marking, both off-line and on-line using eye-tracking in a looking-while-listening paradigm.

4.1. Participants

Fifteen children with DLD (4 girls, mean age 7;2, range 5;3-9;4), and 30 age-matched typically developing children (13 girls, mean age 7;8, range 4;9-9;5), participated. We also tested 30 language-matched controls, but their data will not be presented. Language performance of children with DLD fell 1 standard deviation or more below the mean on at least two subtests of a standardized language tests battery. Their non-verbal intelligence and hearing abilities were in the normal range. Children were recruited at the Babylab of the University of Potsdam or in language therapy centers in Potsdam and Berlin.

4.2. Material

The goal was to test the effect of embedded constituent type, either a lexical NP or a personal pronoun, on the processing of subject and object relatives that were disambiguated by case marking. German subject and object relatives have the same surface word order, allowing the construction of minimal pairs of the two relative clause types. Two variables, RELATIVE CLAUSE (subject / object relatives) and EMBEDDED CONSTITUENT (lexical NP / pronoun) were crossed to form 4 conditions. The relative pronoun is marked in subject relatives (1)-(2) with nominative (*der*) and in object relatives (3)-(4) with accusative (*den*). The case marking on the embedded constituent–in subject relatives, the accusative-marked *den*/*ihn*; in object relatives, the nominative-marked *der*/*er*–also helps to identify the type of relative clause. There were 8 sentences in each condition and 16 fillers (5), resulting in 48 trials in in total, all digitally pre-recorded.

Wo ist der Igel…
Where is the hedgehog…

(1)	**der**	**den** Käfer	fängt?
	that.NOM	the.ACC beetle	catches
	'that is catching the beetle?'		
(2)	**der**	**ihn**	fängt?
	that.NOM	him.ACC	catches
	'that is catching him?'		
(3)	**den**	**der Käfer**	fängt?
	that.ACC	the.NOM beetle	catches
	'that the beetle is catching?'		

(4) **den** **er** fängt?
that.ACC he.NOM catches
'that he is catching?'

(5) mit der Blume?
'with the flower?'

Each sentence was accompanied by a visual scene with two identical animals on the right and left sides of a computer screen and a third different animal in the middle. Each animal was performing some action, for instance catching, on the following animal, except for the last animal (Figure 2). The scenes were animated videos. In the example figure, the catching animals were moving the net up and down and the left hedgehog was moving its head back and forth. The direction of the movement was from right to left in half of the trials and from left to right in the other half, such that the target position was counterbalanced. In the example figure, the right hedgehog is the target referent in subject relatives and the left hedgehog is the target in object relatives.

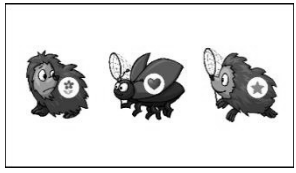

Figure 2. A snapshot from the video accompanying sentences (1)-(5).

4.3. Procedure

Using a looking-while-listening paradigm (Adani & Fritzsche, 2015; Adelt et al., 2017), the experiment was embedded within a searching game. During a familiarization phase, children were instructed to press one of two buttons (left/right) depending on the position of the target animal. At the beginning of each trial, each animal was mentioned and presented individually on the screen, accompanied by a sentence like "Here is a hedgehog, here is another hedgehog, and here is a beetle". The test sentence then followed, whereby the usage of the personal pronoun was felicitous, in that its referent–here, the beetle–could be retrieved from the introductory sentence. The trials were presented via the SMI Experiment Center software using a 22" monitor for stimulus presentation. Eye-movements were recorded via the SMI iView Software using a 60Hz sample rate. The relevant part of the sentence is the one comprising the relative clause. The mean proportion of looks to the target referent within this temporal region was taken as a proxy for accurate processing of the sentence. The button press responses yielded an off-line accuracy measure that will not be presented here.

4.4. Results

The eye movement results are shown in Figure 3. Overall, children look more to the target referent in object relatives than in subject relatives. This pattern reflects a preference of the children to look at the last animal in the scene (the patient). Crucially, this gaze pattern emerges even before the linguistic input is heard, and later children either keep their look on that animal if they think it is the target referent, or they switch to another referent according to how they parse the sentence. That is, this pattern does not appear to depend on the experimental condition. Typically developing children look more to the target in object relatives when the embedded constituent is a pronoun, rather than a lexical NP. The opposite pattern emerges in subject relatives, where an embedded pronoun triggers less target looks than a lexical NP. In children with DLD, there is no effect of the type of embedded constituent.

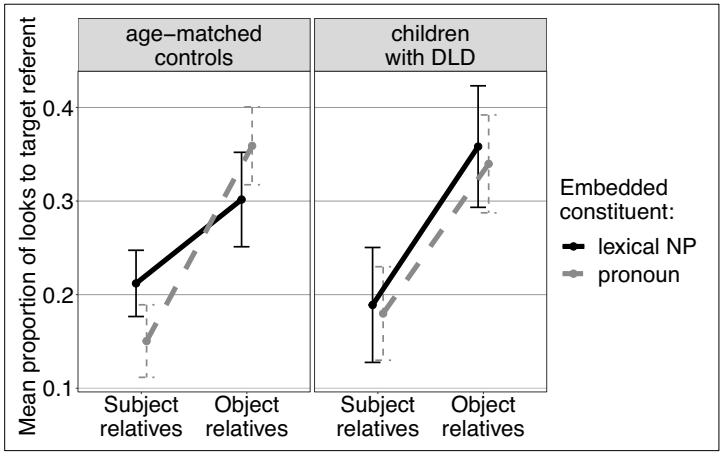

Figure 3. Mean proportion of target looks (with 95% confidence intervals) broken by relative clause type (x-axis) and group (panels). The black solid lines stand for lexical NP; the dashed gray lines stand for pronoun.

4.5. Frequentist analysis

For the analysis we transformed the proportion of target looks into empirical logit, yielding a normally distributed dependent variable that can be used in a weighted linear mixed model (Barr, 2008; Donnelly & Verkuilen, 2017). There were 3 independent variables: RELATIVE CLAUSE (object vs. subject relatives), EMBEDDED CONSTITUENT (lexical NP vs. pronoun) and GROUP (controls vs. DLD). The levels of each of the three factors were coded as 0.5 and -0.5. We first tried to fit a model with all main effects and interactions in the fixed effects part, and with the maximal random effects structure allowed by the design (Barr et al., 2013), using the *lme4* package (Bates et al., 2015). The formula is shown in (6). This model failed to converge, probably because it is too complex for the

relatively small amount of data. We therefore reduced the random effects structure in a piecemeal fashion until we reached a model that converged (7). Here we do not estimate the random intercepts, and therefore also not the correlations between random intercepts and slopes. Moreover, we had to exclude the interaction slope from the random effects part for participants.

(6) elog ~ relative_clause * embedded_constituent * group +
(relative_clause * embedded_constituent | subject_id) +
(relative_clause | item_id)

(7) elog ~ relative_clause * embedded_constituent * group +
(0 + relative_clause + embedded_constituent | subject_id) +
(0 + relative_clause | item_id)

The reduced model is not optimal, since it probably does not reflect the process by which the data were generated in the experiment. Table 1 summarizes the information on the fixed effects parameters. The only significant effect is the main effect of RELATIVE CLAUSE, reflecting the overall increased target looks in object relatives than in subject relatives. But there is no indication for an interaction with EMBEDDED CONSTITUENT and GROUP.

Table 1. Fixed effects parameters in the Frequentist model.

	Estimate	Std. error	t-value
relative_clause	0.44	0.07	6.05
embedded_constituent	0.05	0.07	0.81
Group	-0.01	0.06	-0.18
relative_clause : embedded_constituent	-0.07	0.14	-0.52
relative_clause : group	-0.03	0.13	-0.22
embedded_constituent : group	-0.07	0.14	-0.55
relative_clause : embedded_constituent : group	-0.28	0.26	-1.09

4.6. Bayesian analysis

Next, we will fit the same type of model in a Bayesian framework, using the R package *brms* (Bürkner, 2017, 2018). We will estimate a posterior distribution for each model parameter, and our hypothesis testing will be based on inference from the posteriors of the fixed-effects parameters. The package uses the probabilistic language Stan (Carpenter et al., 2017) to estimate the posteriors, but it does so with the same syntax from the *lme4* package. This facilitates enormously the shift into the Bayesian world for people with *lme4* experience.

The first thing we do, for the sake of inference, is to center all variables around zero. We do this by re-cording the levels of each factor into -1 and 1. For example, for the factor RELATIVE CLAUSE, subject relatives are coded as -1 and object relatives as 1. Thus, for the main effect of relative clause type, a positive

parameter coefficient will reflect more target looks in object than in subject relatives. This coding treats zero as the point of "no difference" between the two relative clause types. The same coding and logic hold for the factors EMBEDDED CONSTITUENT (Pronoun=-1, Lexical NP=1) and GROUP (children with DLD=-1, control children=1). As for the model structure, we fit the model with the maximal random effects structure. Using the proposed Bayesian analysis, fitting complex models to a small amount of data will not result in convergence failure (as long as the model is implemented correctly).

In the next step, we decide about the priors. As the basis for priors, we take 3 previous studies we conducted (Adani et al., under revision; Haendler et al., 2015; Haendler & Adani, unpublished), in which we used similar design, experimental setup and task to test relative clauses with children. The results from these studies will be used to define informative priors for model parameters that concern the factors RELATIVE CLAUSE and EMBEDDED CONSTITUENT, the two factors that were used in the design of the past studies as well as in the current one. For example, for the main effect of EMBEDDED CONSTITUENT the prior is defined as a normal distribution with mean 0.005 and standard deviation 0.07 (on empirical logit scale). This informative prior gives a rather restricted range of outcomes, between -0.2 and 0.2.

By contrast, in the previous studies we only had typically developing children, limiting our ability to make assumptions about the performance of children with DLD in the current experiment. Hence, parameters involving the factor GROUP are assigned weakly informative priors. For instance, the prior for the main effect of GROUP is a normal distribution with mean 0 and standard deviation 7, yielding a weakly informative prior with a large range of possible outcomes, between -20 and 20. A commented R code showing how to assign the priors and implement the model is available under the following link: https://osf.io/rm25h/.

The model output is a large data frame with as many columns as the number of model parameters. Each column is composed of thousands of numbers that compose the posterior distribution for that model parameter. Typically, we are interested in the fixed effects parameters, so we extract from the output data frame the columns that refer to these parameters.

At this point, the posteriors can be plotted (Figure 4) and summarized in different ways. One common way is to draw the 95% credible intervals (CrI), which demarcate 95% of the distribution. We then check whether zero, the point of "no difference", lies outside the 95% CrI. Another way is to calculate the probability of the distribution being smaller or greater than zero. Let us take for instance the main effect of RELATIVE CLAUSE. Zero is excluded from the 95% CrI of this parameter's posterior (in Figure 4, in the top-most posterior, zero is excluded from the light-gray area of the distribution), and the probability that the posterior is greater than zero is close to 1. This result indicates strong evidence that there are more target looks in object relatives (coded as 1) than in subject relatives (coded as -1). Now, let us look at the main effect of GROUP. Zero is

included in the 95% CrI and there is a probability of 0.9 that the posterior lies in the range of negative numbers. Thus, the evidence for this effect is rather weak.

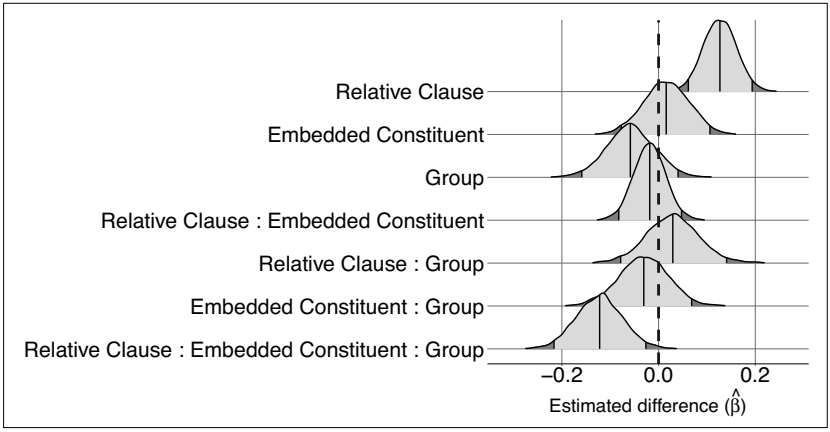

Figure 4. Posterior distributions for the Bayesian model's fixed effects parameters. The estimated difference (x-axis) is on empirical logit scale. Zero is marked with a vertical dashed line. The vertical line in the middle of each distribution is its mean. The two small vertical lines that delimit the light-gray area are the 95% credible intervals.

Note that there are no "significant" or "non-significant" effects. Rather than being binary, inference is formulated in terms of stronger or weaker evidence for a particular effect. More precisely, we talk about the probability that the true parameter of an effect is included in the estimated posterior distribution. But the amount of evidence for a particular effect–whether there is big/small probability that its value is included in the posterior–might be discussed, and argued for or against (for example, when an effect is expected *a priori* to be particularly weak; cf. general discussion in Pozniak et al., 2019).

A last note about the results. If we look at the posterior for the parameter of the three-way interaction of RELATIVE CLAUSE, EMBEDDED CONSTITUENT and GROUP (the last distribution in Figure 4), we see that its 95% CrI exclude zero, the probability that it is smaller than zero being 0.993. Thus, unlike the Frequentist model, the Bayesian model finds evidence for this interaction. This reflects the pattern in the data in which the two embedded constituent types affect subject and object relative clauses differently, a pattern that is different in children with DLD and their typically developing peers.

5. Discussion

The Bayesian analysis we performed has various advantages over the Frequentist approach. First, it allows us to make inference concerning the tested hypothesis given the data. Second, since no Null Hypothesis is involved, there are

no Type I and Type II errors, making the Bayesian approach more suitable for analyzing small data sets (with little data it is preferable to have informative priors, rather than weakly informative ones). Third, Bayesian analysis allows more flexibility in the choice of the model, without risking convergence issues when it is complex. Frequentist models might fail to converge also because parameter estimation is not restricted enough. The incorporation of priors in Bayesian models helps restrict parameter estimation, which generates a model that better fits the data. In this sense, the Bayesian analysis allows us to fit models that we may never be able to fit in the Frequentist framework (Schad et al., submitted).

An additional advantage is the possibility to integrate previous knowledge into the analysis in the form of priors. This reflects the normal course of our scientific research. New studies often follow up on previous ones, and expectations about potential outcomes, based on domain expertise and on previously collected data, can be formulated. Incorporating this knowledge into the analysis, and see how it gets updated with each new data we collect, is a plus provided by the Bayesian approach.

Priors are one aspect of the Bayesian approach that might be perceived by newcomers as problematic or difficult to come about. Priors are defined by the experimenter and it is not always clear which prior should be used. Importantly, different priors might affect the posteriors differently. This raises the question of whether the Bayesian approach allows the experimenter too much subjectivity. We think it does not. First, the Frequentist approach is not less subjective; on the contrary, it is critically concerned by the issue of experimenter degrees of freedom (Roettger, 2019). Moreover, there are several steps that can and should be taken in order to check the appropriateness of the chosen priors and of the model. We did not include these tests here for reasons of space, but more details can be found in a recent tutorial paper (Schad et al., submitted).

Another aspect of the proposed Bayesian analysis that might hold back newcomers is the fact that the result is not a definitive answer (significant or non-significant effect), but rather a probability distribution. However, many scientists anyway adopt a Bayesian way of making statistical inference, when defining non-significant effects as "almost/marginally significant" and the like. Such formulations clearly show that researchers would like to discuss the possibility of having varying degrees of meaningfulness in their data. Moreover, from a purely theoretical perspective, doing statistical inference based on an entire distribution (Bayesian posterior), rather than merely on a point estimate (Frequentist parameter coefficient estimate, standard error and t-/p-value), is an advantage. Ultimately, we want to estimate how far the effect we observe lies with respect to the true parameter of that effect. An outcome that is a probability distribution gives much more information for the purpose of estimating what that true parameter might be, making the picture we get from the data clearer.

This point is important particularly for studies, like the one presented here, where participants belong to a clinical population that presents a specific profile. In such cases, the researcher would like to have clear cut-off points or ranges that are conventionally recognized, and to draw conclusions with consequences for

other patients of the same condition. The greater amount of information that a distribution outcome yields is an advantage also in this case. Furthermore, common conventions, like the 95% credible intervals, are typically defined for particular research domains.

In conclusion, considering the advantages and disadvantages of both analysis methods, the Bayesian analysis emerges as a powerful alternative to the Frequentist approach, especially with small samples. The Bayesian framework is thus an advantageous data analysis method in the field of language development.

References

Adani, Flavia and Fritzsche, Tom (2015). On the relation between implicit and explicit measures of child language development: Evidence from relative clause processing in 4-year-olds and adults. In Grillo, Elizabeth and Jepson, Kyle (Eds.), *Proceedings of the 39th Annual Boston University Conference on Language Development* (pp. 14-26). Somerville, MA: Cascadilla Press.

Adani, Flavia, Haendler, Yair, Lassotta, Romy, Adelt, Anne, Stadie, Nicole and Burchert, Frank (under revision). Asymmetries in the processing of German relative clauses with and without pronouns.

Adelt, Anne, Stadie, Nicole, Lassotta, Romy, Adani, Flavia and Burchert, Frank (2017). Feature dissimilarities in the processing of German relative clauses in aphasia. *Journal of Neurolinguistics 44*, 17-37.

Barr, Dale J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language 59*(4), 457-474.

Barr, Dale J., Levy, Roger, Scheepers, Christoph and Tily, Harry J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language 68*(3), 255-278.

Bates, Douglas, Mächler, Martin, Bolker, Ben and Walker, Steve (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software 67*(1), 1-48.

Bürkner, Paul (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software 80*(1), 1-28.

Bürkner, Paul (2018). Advanced Bayesian multilevel modeling with the R package brms. The R *Journal 10*(1), 395-411.

Carpenter, Bob, Gelman, Andrew, Hoffman, Matthew D., Lee, Daniel, Goodrich, Ben, Betancourt, Michael, Brubaker, Marcus, Guo, Jiqiang, Li, Peter and Riddell Allen (2017). Stan: A probabilistic programming language. *Journal of Statistical Software 76*(1).

Contemori, Carla and Garraffa, Maria (2010). Comparison of modalities in SLI syntax: A study on the comprehension and production of non-canonical sentences. Lingua 120(8), 1940-1955.

Dienes, Zoltan (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science 6*(3), 274-290.

Donnelly, Seamus and Verkuilen, Jay (2017). Empirical logit analysis is not logistic regression. *Journal of Memory and Language 94*, 28-42.

Haendler, Yair, Kliegl, Reinhold and Adani, Flavia (2015). Discourse accessibility constraints in children's processing of object relative clauses. *Frontiers in Psychology 6:860*.

Kimball, Amelia, Shantz, Kailen, Eager, Christopher and Roy, Joseph (2019). Confronting quasi-separation in logistic mixed effects for linguistic data: A Bayesian approach. *Journal of Quantitative Linguistics 26*(3), 231-255.

Kruschke, John K. (2015). *Doing Bayesian data analysis, Second Edition: A tutorial with R, JAGS and Stan*. Burlington, MA: Academic Press / Elsevier.

Kruschke, John K. and Liddell, Torrin M. (2018). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulleting and Review 25*, 178-206.

Matuschek, Hannes, Kliegl, Reinhold, Vasishth, Shravan, Baayen, Harald and Bates, Douglas (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language 94*, 305-315.

McElreath, Richard (2016). *Statistical Rethinking. A Bayesian Course with Examples in R and Stan*. Boca Raton, FL: CRC Press.

Nicenboim, Bruno, Roettger, Timo B. and Vasishth, Shravan (2018). Using meta-analysis for evidence synthesis: The case of incomplete neutralization in German. *Journal of Phonetics 70*, 39-55.

Nicenboim, Bruno and Vasishth, Shravan (2016). Statistical methods for linguistic research: Foundational Ideas – Part II. *Language and Linguistics Compass 10*, 591-613.

Pozniak, Céline, Hemforth, Barbara, Haendler, Yair, Santi, Andrea and Grillo, Nino (2019). Seeing events vs. entities: The processing advantage of Pseudo Relatives over Relative Clauses. *Journal of Memory and Language 107*, 128-151.

Roettger, Timo B. (2019). Researcher degrees of freedom in phonetic research L*aboratory Phonology: Journal of the Association for Laboratory Phonology 10*(1): 1, 1–27.

Schad, Daniel J., Betancourt, Michael and Vasishth, Shravan (submitted). Towards a principled Bayesian workflow: A tutorial for cognitive science.

Sorensen, Tanner, Hohenstein, Sven and Vasishth, Shravan (2016). Bayesian linear mixed models using Stan: A tutorial for psychologists, linguists and cognitive scientists. *The Quantitative Methods for Psychology 12*(3), 175-200.

Stavrakaki, Stavroula (2002). A-bar movement constructions in Greek children with SLI: evidence for deficits in the syntactic component of language. In: Fava, Elisabetta (Ed.), *Clinical Linguistics: Theory and applications in speech pathology and therapy* (pp. 131-153). Amsterdam, NL: John Benjamins Publishing.

Vasishth, Shravan and Nicenboim, Bruno (2016). Statistical methods for linguistic research: Foundational Ideas – Part I. *Language and Linguistics Compass 10/8*, 349-369.

Proceedings of the 44th annual Boston University Conference on Language Development

edited by Megan M. Brown
and Alexandra Kohut

Cascadilla Press Somerville, MA 2020

ISSN 1080-692X
ISBN 978-1-57473-057-9 (2 volume set, paperback)

Ordering information

To order a copy of the proceedings or to place a standing order, contact:

Cascadilla Press, P.O. Box 440355, Somerville, MA 02144, USA
phone: 1-617-776-2370, sales@cascadilla.com, www.cascadilla.com