Received: 26 May 2020 | Revised: 7 January 2021 | Accepted: 22 January 2021

DOI: 10.1111/desc.13097

Developmental Science WILEY

PAPER

Sequential Bayes Factor designs in developmental research: Studies on early word learning

Nivedita Mani[1,2] | Melanie S. Schreiner[1,2,3,4] | Julia Brase[1] | Katrin Köhler[1] | Katrin Strassen[1] | Danilo Postin[1] | Thomas Schultze[2,5]

[1]Psychology of Language Department, University of Göttingen, Göttingen, Germany

[2]Leibniz ScienceCampus "Primate Cognition", Göttingen, Germany

[3]Clinic for Cognitive Neurology, University of Leipzig, Leipzig, Germany

[4]Department of Neurology, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

[5]Department of Economic and Social Psychology, University of Göttingen, Göttingen, Germany

Correspondence
Nivedita Mani, Psychology of Language Department, Georg-Elias-Müller-Institute for Psychology, Gößlerstr. 14, 37073 Göttingen, Germany.
Email: nmani@gwdg.de

Funding information
Leibniz ScienceCampus Primate Cognition

Abstract

Developmental research, like many fields, is plagued by low sample sizes and inconclusive findings. The problem is amplified by the difficulties associated with recruiting infant participants for research as well as the increased variability in infant responses. With sequential testing designs providing a viable alternative to paradigms facing such issues, the current study implemented a Sequential Bayes Factor (SBF) design on three findings in the developmental literature. In particular, using the framework described by Schönbrödt and colleagues (2017), we examined infants' sensitivity to mispronunciations of familiar words, their learning of novel word-object associations from cross-situational learning paradigms, and their assumption of mutual exclusivity in assigning novel labels to novel objects. We tested an initial sample of 20 participants in each study, incrementally increasing sample size by one and computing a Bayes Factor with each additional participant. In one study, we were able to obtain moderate evidence for the alternate hypotheses despite testing less than half the number of participants as in the original study. We did not replicate the findings of the cross-situational learning study. Indeed, the data were five times more likely under the null hypothesis, allowing us to conclude that infants did not recognize the trained word-object associations presented in the task. We discuss these findings in light of the advantages and disadvantages of using a SBF design in developmental research while also providing researchers with an account of how we implemented this design across multiple studies.

KEYWORDS
cross-situational learning, effective research, informative research, mispronunciation task, mutual exclusivity, Sequential Bayes Factor

1 | INTRODUCTION

Developmental research is resource-, personnel- and time-intensive. Researchers spend time and money establishing contact and relationships of trust with caregivers, who already have demands on their time and may not be able to commit to participating in research to the same extent as undergraduate students. Even when a sample of participants has been recruited for a study, the above average drop-out rates of such studies lead to many testing sessions yielding little actual data for the study (Miller, 2012). As a consequence

perhaps, average sample sizes in developmental research are considerably lower (Oakes, 2017) than what is typical in research with more accessible participant pools. Smaller sample sizes, in turn, lead to studies not having sufficient power to detect the effects under investigation and inconclusive or spurious findings driven by isolated participants. Issues of power are further amplified in developmental research due to the fact that studies with infants need to be shorter to maintain their attention throughout, leading to fewer trials per condition and, consequently, greater variability in infant performance. Especially given the time and effort typically invested in recruiting and testing younger participants, inconclusive or spurious findings—the infamous $p = 0.08$ (Schönbrodt et al., 2017)—are frustrating to say the least and deterring at worst. Against this background, the current study examines the practicability of using Sequential Bayes Factor design (SBF design, see Schönbrodt & Wagenmakers, 2018; Schönbrodt et al., 2017; Stefan et al., 2019 for easily accessible reviews) in developmental research as a more "informative and efficient" (Schönbrodt & Wagenmakers, 2018) alternative to other more commonly used design and analytic procedures.

Broadly speaking, following an SBF design (Schönbrodt et al., 2017), researchers collect data from an initial pre-specified sample of participants and infer the probability that H_1 is true (an effect exists in a given population or between populations) relative to the probability that H_0 is true (the effect does not exist) given the observed data. This ratio, called the *posterior odds*, tells us how much more likely one hypothesis is relative to the other. For example, posterior odds of 10 tell us that, given the data, H_1 is 10 times more likely to be true than H_0. If the posterior odds exceed a pre-specified minimum level of evidence for the alternative or null hypothesis, known as the H_1 boundary or the H_0 boundary, then the researcher can choose to stop testing and draw their conclusions based on this initial sample. Critically, if the posterior odds do not cross the pre-specified threshold, the researcher can continue testing participants until either the threshold for H_1 or H_0 is crossed. Once either threshold is crossed, the researcher can stop testing and draw their conclusions regarding the strength of the evidence for either hypothesis.

We can compute the posterior odds as follows:

$$\frac{P(H_1|D)}{P(H_0|D)} = \frac{P(D|H_1)}{P(D|H_0)} \times \frac{P(H_1)}{P(H_0)}.$$

Here, $P(H_1)$ and $P(H_0)$ are the prior probabilities we assign to H_1 and H_0, and their ratio is called the *prior odds*. The prior odds indicate how probable we believe H_1 to be relative to H_0 before we have had a look at the data. Finally, $P(D|H_1)$ and $P(D|H_0)$ are the likelihood of obtaining the observed data if either H_1 or H_0 were true. Their ratio is called the Bayes Factor (BF). The BF quantifies the extent to which the data are more probable were H_1 to be true relative to when H_0 were true. It serves as an updating factor that tells us how we must change our prior odds in the light of the data. Since we can only compute the BF, but are ultimately interested in the posterior odds when making statements about the existence or absence of a particular effect, it is common to assume, for the sake of convenience, that H_1 and H_0 are equally probable as priors, that is, to set the prior odds to 1 (Schönbrodt et al., 2017). In this case, the BF equals the posterior odds, and we can interpret the relative likelihood of the data given both hypotheses as their relative probability of being true. For the purpose of this paper, we will only consider the case of equal prior odds and will, thus, treat the BF as the posterior odds, that is the probability that H_1 is true relative to the probability that H_0 is true, given the observed data.[1]

Research Highlights

- A Sequential Bayes Factor (SBF) design replication of three key developmental findings.
- We tested infant sensitivity to mispronunciations of words, cross situational word-object association learning and mutual exclusivity based learning.
- We replicate sensitivity to mispronunciations and mutual exclusivity based learning but not cross-situational word-object association learning.
- SBF designs allow greater flexibility in developmental research allowing further testing in cases of inconclusive findings.
- Advantages and issues associated with SBF designs.

Following guidelines for accepted levels of evidence (Lee & Wagenmakers, 2013), BFs between 1 and 3 are considered *anecdotal evidence*, BFs between 3 and 10 *moderate evidence*, BFs between 10 and 30 *strong evidence* and BFs over 30 *very strong evidence* for H_1. The inverse is generally true for H_0 (1/3 to 1—*anecdotal evidence*, 1/10 to 1/3—*moderate evidence*, 1/30 to 1/10—*strong evidence* and under 1/30—*very strong evidence*). Moderate evidence is considered the lowest acceptable evidence threshold for accepting either H_0 or H_1. While at first glance, a BF of 3 (or 1/3) might not seem overwhelming, even this relatively low evidence threshold is more conservative than using a p value of 0.05 (Wetzels et al., 2011). While most researchers use symmetric thresholds when testing hypotheses using BFs (e.g., they decide to accept H_1 if BF >3 and to accept H_0 if BF <1/3), it is permissible to set an asymmetric threshold for H_0, given that evidence for H_0 accumulates at a much slower pace (Jeffreys, 1961; Schönbrodt & Wagenmakers, 2018, p. 133).

The critical difference to designs employing standard null hypothesis significance testing (NHST) is that the researcher is not constrained by a pre-specified sample size whereupon they have to stop data collection. Rather, following the initial computation of BF, the researcher can either choose to stop data collection or continue collecting data (referred to as *optional stopping*) until they are convinced they have accumulated reliable evidence for either hypothesis. Such optional stopping or continuation of data collection after testing the last planned participant (and the repeated statistical significance testing that comes along with it) is not possible in the NHST paradigm due to issues of inflation of Type 1 error rates (Armitage et al., 1969). In contrast, in Bayesian hypothesis testing,

the pre-defined criterion for the BF can be chosen such that there is a high probability that evidence for the wrong hypothesis cannot be obtained (Edwards et al., 1963).[2] Note that the initial sample size chosen in SBF designs for the first computation of the BF is a combination between a convenience sample (how many people can I reasonably test before I begin to compute sequential BF?) and a judgement of an adequate sample size that would reliably provide evidence of a given effect (how many people do I need to test before I can believe the evidence?). This is in contrast to justifications of sample sizes that are either based on analyses of the power to detect the effect under investigation or examples of prior studies on similar topics. SBF can also accommodate researchers who prefer to think about statistical inference in terms of type I and type II errors using a Bayes Factor design analysis (BFDA; Schönbrodt & Wagenmakers, 2018). This entails a simulation that returns the expected type I and/or type II error rates depending on the planned sample (this could either be a fixed sample size or, in the case of SBF, the planned minimal, incremental, or maximal sample size), the expected effect size, and the aspired BF thresholds. We include a design analysis for our own studies in Appendix A.

Note that a researcher can, in principle, choose to stop testing before a pre-specified inference threshold is crossed for monetary or organizational reasons. For instance, evidence for H_0 typically accumulates at a much slower pace which may necessitate testing to stop before the threshold for H_0 is reached in some cases (Schönbrodt & Wagenmakers, 2018). Similarly, a researcher may choose to continue testing even after a particular threshold is crossed to increase their confidence in the stability of the effect given potential issues of false positive or false negative evidence. Such subjective decisions with regards to optional stopping will obviously have implications for the researchers' and readers' confidence in their findings, therefore allowing researchers to stop testing earlier and reporting their reduced confidence in their findings or continuing testing until they are more confident in the stability of the effect or other factors intervene.

There are a number of obvious advantages as well as some disadvantages to SBF designs that are highlighted by Schönbrodt et al. (2017). With a focus on the additional difficulties associated with recruiting and testing infants, the flexibility with regards to optional stopping is a clear win. The authors note that SBF designs may be between 50% to 70% more efficient, allowing researchers to collect data from fewer infant participants when the effect is strong and to continue collecting data when the effect is weak.

Further flexibility is afforded by the fact that the stopping rule is "a suggestion, not a prescription" (Schönbrodt et al., 2017, p. 14). Thus, if it is important to a researcher to judge the strength or stability of an effect, they can continue testing even after the threshold is crossed for information on long-term rates of false evidence. At the same time, the possibility of incremental testing until either threshold is crossed allows the researcher more definitive evidence for or against an effect, so that resources are not unnecessarily wasted testing a pre-specified sample of infants to end up with an inconclusive $p = 0.06$.

While Schönbrodt et al. (2017) also go into considerable detail about the disadvantages of such a design, we will focus here on one of the decisions that the researcher needs to take regarding the prior beliefs that they incorporate into the analyses. The Bayesian world is divided into objective Bayesians who recommend default prior distributions, for example, the recommendation made in Schönbrodt et al. (2017) for a Cauchy prior distribution (JZS prior, Schönbrodt et al., 2017), and subjective Bayesians, who recommend incorporating information of the effect sizes reported in previous studies into the analyses. The argument for defining subjective priors is they reflect a researcher's true belief in the size of an effect before they analyze their data. In some sense, using subjective priors aligns best with the core idea of Bayesian reasoning. While objective priors do not necessarily capture a researcher's true a-priori beliefs about the effect of interest, they help to navigate the problem that different researchers may hold different a-priori beliefs and, thus, also different posterior beliefs about an effect. A second important argument for objective priors is specific to using BFs as the inference criterion. As Gronau et al. (2020) point out, priors should meet two important criteria. The first is predictive matching, that is, if the data are completely uninformative, the BF should be 1. The second criterion, information consistency, states that as the evidence for H_1 increases (to infinity) so should the BF. While some objective priors, such as those implemented as defaults for Bayesian *t*-tests, meet these criteria, subjective priors usually do not.[3] Given this debate, we note that, following the recommendation made in Schönbrodt et al. (2017), we chose to report a sensitivity analysis here to examine how the strength of evidence for or against the effect varied across a range of priors.

1.1 | The current study

Here, we test three oft-reported effects in the developmental literature using an SBF design (see Appendix B for details of studies replicating these effects, reported sample sizes and findings as well as BFs that we computed for each of these studies based on the details provided in the individual manuscripts). First, we examined children's sensitivity to mispronunciations of familiar words in an eye-tracking task (Swingley & Aslin, 2000). Children were presented with images of familiar objects side-by-side on a screen and heard either a correct or incorrect pronunciation of the label for one of these objects, for example, *baby* mispronounced as *vaby*. The finding that children looked longer at the target when cued by a correct pronunciation relative to a mispronunciation of its label suggests that children were sensitive to mispronunciations of the words and represent words with adequate phonological detail even early in development.

Next, we examined children's learning of novel word-object associations via cross-situational statistics (Smith & Yu, 2008). Here, children were presented with trials where they saw two novel objects side-by-side on the screen and heard two novel labels without being told which of the two objects each of the labels referred to.

Across multiple such trials, children can infer the intended label-object mappings based on the fact that every occurrence of a particular object is accompanied by a particular label. Children were then presented with two of the previously trained objects and the label for one of these objects. Children looked longer at the correct object when presented with the label for this object, suggesting that they can use cross-situational statistics to infer the intended referents of ambiguous words.

Finally, we examined children's assumption of mutual exclusivity in assigning labels to objects (Markman & Wachtel, 1988). According to the mutual exclusivity bias, children refrain from assigning an unfamiliar label to an object whose label they are familiar with. Rather, they ought to assign the unfamiliar label to an object whose label is unknown, that is, to an unfamiliar object. Here, children were presented with two objects, only one of which they already knew the label of, for example, a banana (familiar) and a cherry pitter (unfamiliar). Half of the children were assigned to the *Novel label condition* where they were asked to "Show me the x," where x was an unfamiliar word, while the other half were assigned to the *Control condition* where they were asked to "Show me one." Children chose the unfamiliar object, for example, the cherry pitter, significantly above chance only in the *Novel label condition* but not in the *Control condition*, suggesting that they responded in keeping with a mutual exclusivity bias.

We tested these hypotheses using default Bayesian *t*-tests as implemented in the *BayesFactor* package for R (Rouder et al., 2009). In the default settings, H_0 is a point hypothesis, that is, under H_0, the only possible value for the effect size *d* is zero, and the whole probability mass is concentrated at an effect size of zero. H_1 entails all non-zero effect sizes but, as opposed to frequentist statistics, not all possible values of the effect size are deemed equally probable a priori. Instead, smaller effect sizes are a priori considered more probable, and this default belief is modeled using an objective Cauchy prior on the effect size *d* with the default scaling factor of $\sqrt{2}/2$. The Cauchy distribution is a t-distribution with 1 *df*. It has a fixed shape and includes no parameters that would influence its shape. However, one can use the scaling factor to make the distribution wider/flatter or narrower/taller (greater values lead to wider and flatter distributions). The prior distribution of the effect size is the only difference between H_1 and H_0 in the default Bayesian *t*-test, that is, all other parameters such as the variance have the same priors. Therefore, computing the BF in a Bayesian *t*-test tells us how likely it is to obtain the observed data given that the effect is zero (H_0) or that any of the non-zero effect sizes is true (H_1), all other things being equal. Figure 1 illustrates how the prior probability for the effect size looks under H_0 and under H_1 in the default Bayesian *t*-test.

Across all studies, we used an SBF design in keeping with the recommendations of Schönbrodt et al. (2017). The first recommendation regards the choice of initial sample size where we, in keeping with Schönbrodt et al. (2017), chose an initial sample of 20 participants in each study and subsequently incrementally tested participants (adding one participant at a time). We set our threshold for optional stopping at 1/5 < BF < 5. In other words, were the BF to drop to <1/5 or >5, we would stop testing. Schönbrodt et al. (2017) recommend this setting due to the lower rates of false positive and false negative evidence with this boundary. While we follow Schönbrodt et al's recommendations in the current study, we direct the reader to Schönbrodt et al. (2017) for details of expected stopping-n across a range of SBF designs with different expected effect sizes and Bayes thresholds, which could be used to plan initial sample sizes and Bayes thresholds based on the expected effect sizes of planned studies. Thus, the researcher may want to plan a larger initial sample size for a study with a smaller expected effect size and a conservative BF threshold, or plan a smaller initial sample size (as chosen here) or less conservative Bayes threshold when replicating a known effect with a larger effect size.

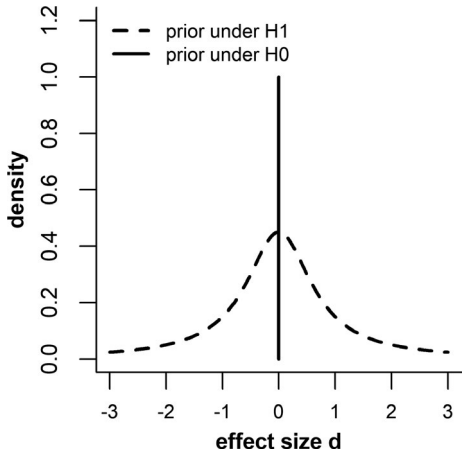

FIGURE 1 Prior probabilities of different values of the effect size *d* in the default Bayesian *t*-test of the *BayesFactor* package for R. The prior probability under H_1 is a Cauchy distribution with the default scaling factor of $\sqrt{2}/2$

2 | METHOD

We will report all methodological details for each replication individually, beginning with the mispronunciation sensitivity replication (Swingley & Aslin, 2000), followed by the cross-situational learning (Smith & Yu, 2008), and finally the mutual exclusivity replication (Markman & Wachtel, 1988).

2.1 | Mispronunciation sensitivity task

2.1.1 | Participants

We recruited 32 children for this study. 21 children aged between 18 and 23 months (*M* = 21.14 m, range 18–23.24 m, 8 female) were included in the final analyses (Swingley & Aslin, 2000). Three bilingual children were tested as pilot children and were excluded from the final analyses. In addition, the data from six children had to be excluded from the analyses due to technical issues caused by migration

of the laboratory to a new room prior to starting the experiment. Two additional children were invited to testing but did not want to participate upon arrival.

All children tested across all three studies reported here were recruited from the laboratory database. All children were born full-term, had normal hearing and vision, and were reported to be monolingual German learners. Children received a book in return for their participation in the study. The study received ethical approval from the Ethics committee of the Psychology department of the University (Number 190b).

2.1.2 | Stimuli

Six colorful images of objects known to be familiar to 18-month-olds were chosen as the critical test stimuli for the study (apple, digger, ball, cookie, star, and teddy). The images were yoked together in pairs so that each image was always presented together with the other image in the pair. Each image, embedded in a gray background, spanned 480 × 390 pixels and appeared to the left and right of the monitor separated by 400 pixels (see Figure 2). In addition, we presented filler trials containing images of four familiar objects (caterpillar, duck, car, and shoe) against a gray background. The images in filler trials were presented in the four quadrants of the screen (counterbalanced across trials) equidistant from the center of the screen.

A female native speaker of German recorded pairs of sentences containing the critical target words, either correctly or incorrectly pronounced. The first sentence was "Where is the X?" [Wo ist der X], where X was the label of the intended target, either correctly or incorrectly pronounced, followed by a further neutral sentence (e.g., "Can you find it?"/"Do you recognize it?"/"Do you see it?"). The mispronunciations of the words were as follows—*Apfel–Opfel*, *Keks–Teks*, *Stern–Storn*, *Bagger–Dagger*, *Ball–Gall*, *Teddy–Beddy*.

2.1.3 | Procedure

The experiment took place in an eye-tracking booth, where the child sat facing a 40 in screen (1920 × 1080) on which the stimuli were presented using Tobii Pro Studio (version 3.4). Participants were seated either on their caregiver's lap or in a car seat approximately 60 cm away from the screen. The auditory stimuli were presented via loudspeakers located above the screen to the left and right side of the screen. Eye movements and pupil diameter were captured using a Tobii X3-120 eye tracker with a gaze sampling rate of 120 Hz. We ran a 5-point calibration and the experiment only started when at least four of the five points were correctly calibrated.

Children were presented with 28 trials (24 test trials and 4 filler trials) across four blocks, where the order of the trials and the appearance of the target to the left and right side of the screen was counterbalanced across blocks. Each block included six test trials and one filler trial. Each image was the target four times (twice correct, twice mispronounced) and the distractor four times and appeared to the left and right side of the screen an equal number of times. Each target was labelled only once in each block, either with a correct or an incorrect pronunciation.

Each trial began with the presentation of both images in silence. The first sentence began 3 s into the trial and the trial ended 6 s after the onset of the first sentence. Filler trials were presented in silence. Trials were separated by a black screen with a white cross in the middle and the next trial was only initiated when the participants were fixating the cross. The experiment took around 5 min to complete.

2.1.4 | Preprocessing

We aggregated gaze data into 40 ms bins offline and coded for whether the child was fixating the target or the distracter for each of the 40 ms bins. Area of Interests (AOIs) for the images were based on the dimensions of the image including a 50 pixel frame around the image. Data points where one or two of the eyes could not be tracked reliably (validity <2 on Tobii scale) and trials where more than 80% of the data could not be tracked were rejected (cf. Ackermann et al., 2020). Only fixations were retained for analysis (looks at a particular AOI more than 60 ms). We calculated the proportion of target looking (PTL) as the total amount of time infants spent looking at the target divided by the total amount of time spent looking at the target and the distractor. The PTL was calculated individually for each trial including all fixations that took place between 360 ms after the onset of the target word in the trial until 2000 ms after the onset of the target word. The 360 ms cut-off ensured that we only considered eye movements that could reasonably be interpreted as a response to the auditory stimulus (Swingley et al., 1999).

2.2 | *Cross-situational learning task*

2.2.1 | Participants

Fifty-five children were recruited for this study. 43 children aged between 12 and 15 months (*M* = 12.77 m, range 11.93–14.22 m,

FIGURE 2 Images of objects used in the mispronunciation task

21 female) were included in the final analyses (Smith & Yu, 2008). Four bilingual children were tested as pilot children and excluded from the final analyses. In addition, the data from five children had to be excluded from the analyses due to technical issues caused by migration of the laboratory to a new room prior to starting the experiment. Three additional children were tested but excluded from further analyses for not providing data for at least two trials in the test phase.

2.2.2 | Stimuli

Six images of objects (see Figure 3) unfamiliar to children were chosen as novel objects for the current study such that each object was easily discriminable from the other objects and was of roughly the same size on screen. Each image embedded in a gray background spanned roughly 350 × 600 pixels and appeared to the left and right of the monitor. In addition, we presented attention getter trials containing videos between trials to maintain children's interest in the experiment. These consisted, for, for example, of a rotating windmill, a parrot on a cycle, a Teletubby.

In addition, six bisyllabic non-words in keeping with the phonotactics of German "Akan," "Upos," "Basa," "Modi," "Sibu," and "Isot" were chosen as the labels for the novel images. A female native speaker of German recorded isolated tokens of the individual words which were then spliced together differently for training and test trials.

2.2.3 | Procedure

The experiment took place in an eye-tracking booth identical to that of the previous study. Children were assigned to one of six lists which counterbalanced, across children, the word-object pairings that were presented such that each object was labelled with different words across the six lists. In each list, children were presented with 30 training trials and 12 test trials. In each training trial, children were presented with two of the six objects to the left and right side of the screen and heard the two labels that had been assigned to both these objects in a pseudorandomized order. There were no cues provided within each trial as to the word-object assignment. Each training trial lasted 4 s. The onset of the first label in a training trial was at 500 ms, while the second label began 500 ms after the offset of the first label. During training, children were presented with the correct label-object pairing 10 times. Each object appeared five times to the left and five times to the right of the screen in a pseudorandomized order. We ensured that we never presented the same object across two immediately successive trials. Attention-getter trials were interspersed with training trials after the 2nd, 8th, 11th, 18th, 22nd, 26th, and 30th trial.

During test trials, children were presented with two images of two objects side-by-side on screen and heard the label of one of these objects, which was repeated three further times during the trial. The onset of the first label was at 500 ms with at least a 500 ms pause between the offset of the first token and the onset of the second token. Each trial lasted 8 s. Across trials, each object appeared as target once on the left and once on the right hand side of the screen. The order of trials was pseudorandomized so that each object was the target object at least once before it was labelled a second time. Attention-getter trials were interspersed after each test trial. Both test and training trials were only initiated once children fixated a white fixation cross in the middle of the screen.

2.2.4 | Preprocessing

We aggregated gaze data into 40 ms bins offline and coded for whether the child was fixating the target or the distracter for each of the 40 ms bins. Following Smith and Yu (2008; who used manual coding of infant videos), we assigned looks to the left and right hand side of the screen with a 80 pixel gap between the left and right AOI, such that each AOI was 600 × 700 pixels. All other preprocessing steps were identical to the first study. As in Smith and Yu (2008), we calculated the total amount of looking time at the target and distractor for each trial. This was calculated individually for each trial including all fixations that took place between 360 ms after the onset of the target word in the trial until the end of the trial.

2.3 | Mutual exclusivity task

2.3.1 | Participants

Twenty-two children were recruited for this study (Markman & Wachtel, 1988). A sample of 20 children aged between 3 and 4 years (M = 45.5 m, range 40.39–52.06 m) were included in the final analyses. One child was tested as pilot child and was excluded from the final analyses. Another child was tested but excluded from the dataset due to an experimenter error. Children were assigned to a familiar label and a control condition in a pseudorandomized order while controlling for age and sex of the children.

FIGURE 3 Images of objects used in the cross-situational learning task

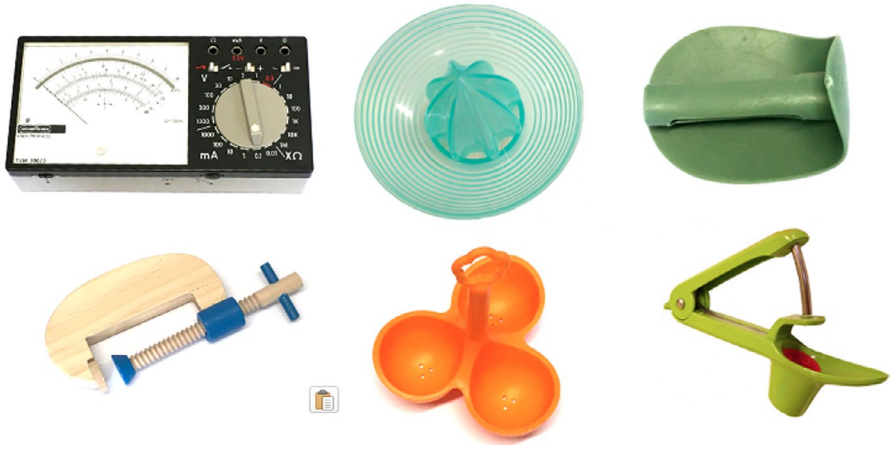

FIGURE 4 Unfamiliar objects used in the mutual exclusivity task

2.3.2 | Stimuli

Six familiar (cup, cow, saw, banana, plate, spoon) and six unfamiliar objects (see Figure 4) were chosen as critical stimuli for the study. In addition, we selected six phonotactically legal German pseudowords, "Mido," "Grasch," "Toma," "Bex," "Dofu," and "Nohle" to be used as labels for the novel label condition.

2.3.3 | Procedure

Each child was tested individually with the experimenter and the child sitting across from each other at a table. The experimenter introduced the child to a frog hand-puppet and asked children to point the frog to the things it was asking for. Importantly, children were told that there was no right or wrong answer. In the novel label condition, children were presented with a familiar and a novel object and asked to identify the referent of the novel label. In the control condition, children were asked to show one of the two objects.

The objects were presented on identical trays located equidistant from the child. For each condition, we created two lists with different pairings of objects and order of presentation of object pairs such that an equal number of children in both conditions saw objects paired together in exactly the same way and presented in the same order. Objects were presented side by side with paired objects being roughly the same size in each pair. The location of the novel object to the left or right side of the child was counterbalanced across trials within each child. We also controlled for the pseudowords that were presented with each pair, with the ordering of pseudowords in one list being the exact opposite of the ordering of pseudowords in the other list. Children were randomly assigned to one of the two lists, ensuring that an equal number of children in both conditions saw each list. The only difference between conditions was that children were asked to find the referent of a novel label in the novel label condition, for example, "Show me the Mido!", and to select one of the objects in the control condition, for example, "Show me one!". Importantly, we always asked children before the objects were placed on the table to ensure that they were attending to the question and to ensure that children did not reach for an object before hearing what they were asked for.

2.3.4 | Preprocessing

We coded the number of times that children chose the unfamiliar object separately for when they were asked to identify either the referent of a novel label or to select one of the two objects. As in the original study, we then compared the number of times (out of a total of six times) that children chose the unfamiliar object in the novel label and the control condition.

3 | RESULTS

We preprocessed and analyzed the data using R (R Core Team, 2020). We analyzed the data using Bayesian *t*-tests as implemented in the *BayesFactor* package (Rouder et al., 2009). For the mispronunciation sensitivity task and the cross-situational learning task, we used paired-sample *t*-tests, whereas for the mutual exclusivity task, we used an independent samples *t*-test. The starting sample size was 20 participants (10 per condition in the mutual exclusivity task). We ran the first test after gathering the initial sample and stopped data collection if the BF exceeded the threshold for accepting H_1 (BF >5) or fell below the threshold for accepting H_0 (BF <1/5). If the result was still inconclusive, we planned to test one additional participant and to recompute the *t*-test until the result was conclusive (i.e., the BF exceeded 5 or fell below 1/5).

In case of the mispronunciation and mutual exclusivity tasks, the data were informative after collecting the initial sample. We note that we mistakenly tested 21 children in the initial sample in the mispronunciation task and therefore report the data with this sample of 21 children. In contrast, we tested 43 children in

the cross-situational learning task. Technically, our final sample size for this task would have been 32 children because the BF fell below 1/5 after testing the 32nd child. However, given the absence of evidence for the H_1, we continued testing to guard ourselves against a false negative. We stopped data collection at 43 participants due to the researcher in charge finishing her project and the BF remaining below 1/5. For each analysis, we not only report the results of the *t*-test but also a point estimate of the effect size *d* as well as the corresponding 95% highest density interval (HDI). The HDI is based on the posterior distribution of the effect size and includes those 95% of the possible values for *d* that are the most probable given the data.

3.1 | Mispronunciation sensitivity task

As shown in Figure 5 (upper left panel), children looked longer at the target in correctly pronounced trials (M = 0.72, SD = 0.09) relative to mispronounced trials (M = 0.65, SD = 0.08), $t(20)$ = 3.10, BF = 8.13, d = 0.61, 95% HDI [0.14, 1.07] (see Figure 6, upper right panel; Swingley & Aslin, 2000).[4] As the BF of 8.13 indicates, H_1 is about eight times more likely to be true than H_0 given the data. Our robustness analysis shows that the evidence in favor of H_1 is robust across a wide range of possible values for the scaling factor of the JZS prior in the effect size (see Figure 5, lower panel).[5]

3.2 | *Cross-situational learning task*

We found no reliable differences in children's looking time to the target (M = 2781 ms, SD = 791) and to the distractor (M = 2709 ms, SD = 718), $t(42)$ = 0.43, BF = 0.18, d = 0.06, 95% HDI [−0.22, 0.35] (see also Figure 6, upper left panel; Smith & Yu, 2008). The BF indicates that H_0 is at least five times more likely than H_1 given the data, and effect size of zero well within the 95% HDI (see Figure 6, upper right panel). The bottom left panel of Figure 6 shows the trajectory of the SBF following collection of the initial sample (n = 20). A robustness analysis showed that support for H_0 is robust when the Cauchy prior on the effect size is wider than the default prior we used. As the prior becomes narrower, the evidence in favor of H_0 weakens. This is to be expected, though, because as the scaling factor r decreases, the Cauchy prior approximates a point distribution at zero (i.e., it approximates the point null hypothesis).[6]

3.3 | Mutual exclusivity task

As shown in Figure 7 (upper left panel), children chose the unfamiliar object more often when asked to identify the referent of a novel label (M = 5.1, SD = 1.10) compared to when they were asked to select one of the two objects (M = 2.2, SD = 1.68), $t(18)$ = 4.55, BF = 93.34, d = 1.75, 95% HDI [0.62; 2.89] (see also Figure 7, upper right panel; Markman & Wachtel, 1988). The data show clear support

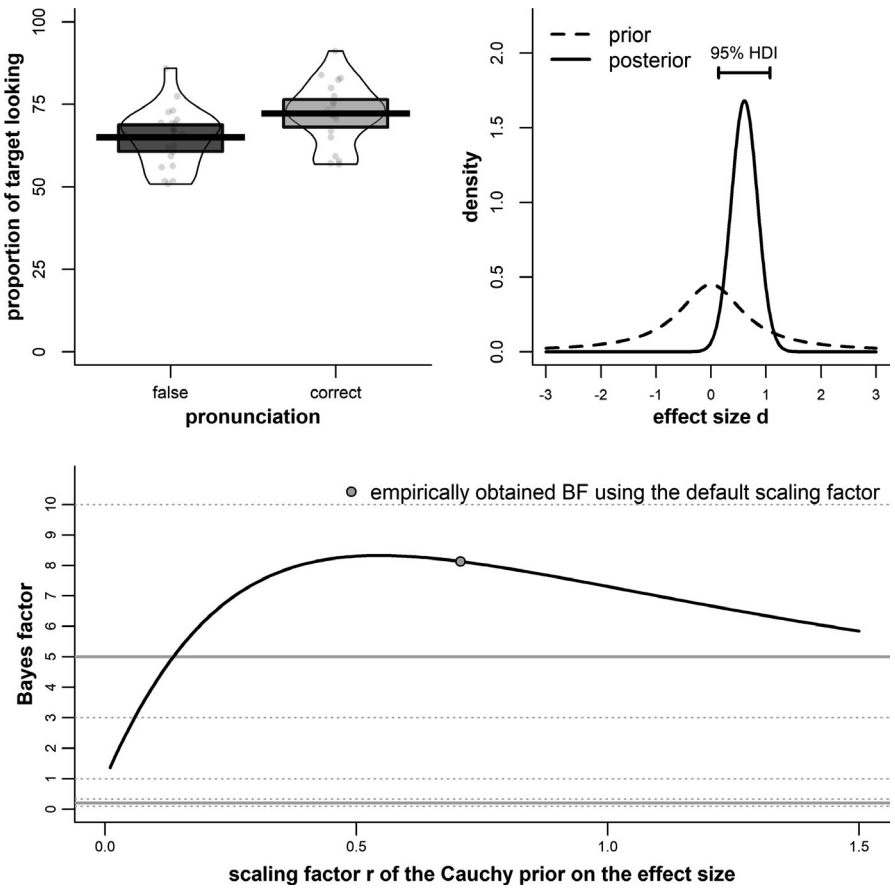

FIGURE 5 Results of the sequential Bayesian analysis of the mispronunciation effect. The upper left panel shows the proportion of time looking at the target by pronunciation condition. The bold horizontal lines represent the mean, while the rectangular boxes denote the 95% highest density interval (HDI) around the mean. The width of the beans is in an indicator of the density with the dots representing individual data points. The upper right panel shows the prior and posterior distribution of the effect size *d* along with the posterior 95% HDI. The lower panel shows the results of the robustness analysis. The Bayes Factor (BF) is plotted as a function of the scaling factor of the prior on the effect size. Dashed gray lines mark different conventional evidence thresholds for accepting H_1 and H_0 while the bold gray lines represent the evidence thresholds of 5 and 1/5 we chose for our analysis

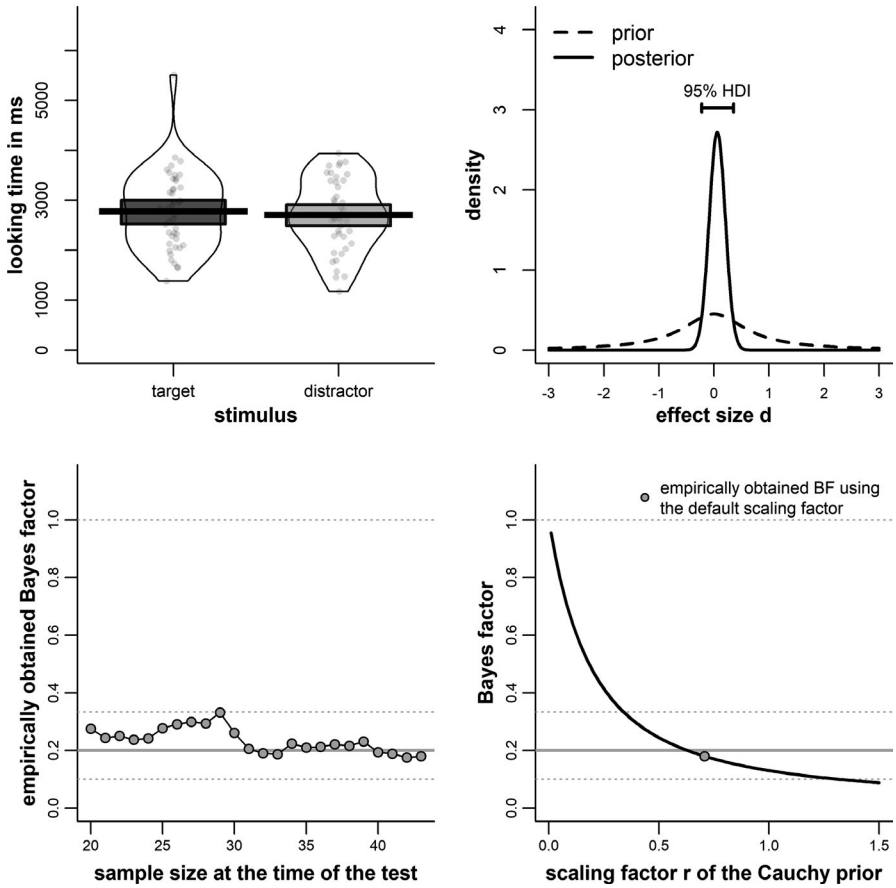

FIGURE 6 Results of the sequential Bayesian analysis of the cross-situational learning task. The upper left panel shows the looking time by type of image. The bold horizontal lines represent the mean, while the rectangular boxes denote the 95% highest density interval (HDI) around the mean. The width of the beans is in an indicator of the density with the dots representing individual data points. The upper right panel shows the prior and posterior distribution of the effect size *d* along with the posterior 95% HDI. The lower left panel shows the trajectory of the Bayes Factor (BF) in the sequential BF analysis. The bold gray line denotes the inference threshold for accepting H_0 (BF = 1/5). The lower right panel shows the results of the robustness analysis. The BF is plotted as a function of the scaling factor of the prior on the effect size. Dashed gray lines mark different conventional evidence thresholds for accepting H_1 and H_0 while the bold gray lines represent the evidence thresholds of 5 and 1/5 we chose for our analysis

for a strong mutual exclusivity effect with H_1 being almost 100 times more likely than H_0. Not surprisingly, the support for H_1 is highly robust even for very narrow priors on the effect size (see the lower panel of Figure 7).[7]

4 | DISCUSSION

Stefan et al. (2019) highlight the need for a balance between informativeness and efficiency in empirical research. As researchers, we aim to set up our studies in such a way as to obtain the maximum amount of information regarding the truth of the hypotheses under consideration whilst using our resources as efficiently as possible. Given that such studies are often financed by the tax payer or other charitable associations and that psychological research with human participants, especially with young infants and children, requires other individuals to invest time and effort in our research, the pursuit of informativeness and efficiency in study design ought to feature more in discussions of experiment design. Our findings contain demonstrations of both superior efficiency and informativeness of an SBF design.

We first address the issue of efficiency using our replication of the Swingley and Aslin (2000) study. The data were seven times more plausible under the alternative hypothesis compared to the null hypothesis, after testing only 21 participants. Compared to the 53 participants tested in the original study, this is a clear win in terms of efficiency. Since the BF of the default Bayesian *t*-test meanders towards infinity were the alternative hypothesis to be true (*consistency*, Morey & Rouder, 2011; Rouder et al., 2012), testing more children should yield BFs indicating stronger support of the hypothesis that children are sensitive to mispronunciations of familiar words. Were the hypothesis under consideration to be more critical to the purposes of the study, researchers would be free to either set a higher minimum sample size or to start off with a more conservative BF threshold from the beginning, or indeed, as we did in one of the studies reported here, continue testing until they received stronger evidence. Thus, the researcher has multiple avenues for flexibility in this approach, with regards to variation in initial sample size, or chosen Bayesian thresholds or indeed to continue testing past the threshold, depending on the expected effect sizes, the novelty of the paradigm, resources available and the ethical ramifications of potentially unnecessary testing. We could also imagine benefits in scenarios common to developmental research where the first in a series of experiments aims to replicate a prior finding, while subsequent studies manipulate the primary finding. In this case, a researcher could set different BF criteria for different experiments, with the caveat that relative interpretation of the effects ought to be constrained by the strength of evidence obtained.

The results of the second experiment speak to the notion of informativeness. The original study (Smith & Yu, 2008) found that 12- to 14-month-olds were able to learn one-to-one word-object

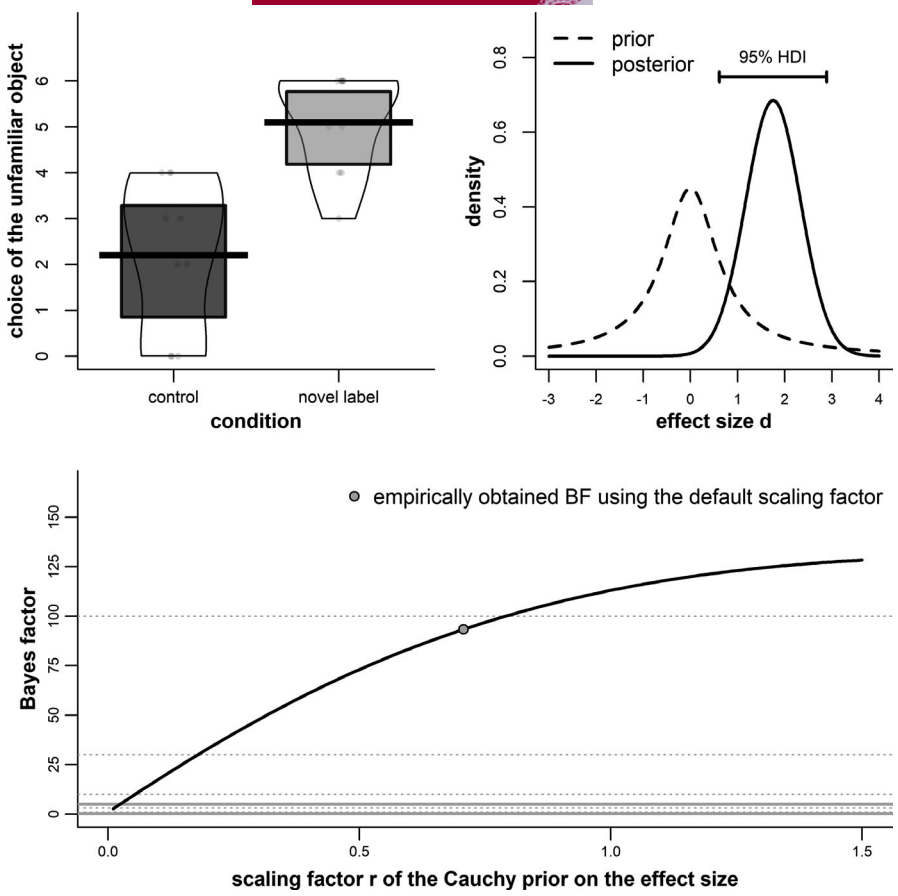

FIGURE 7 Results of the sequential Bayesian analysis of the mutual exclusivity task. The upper left panel shows the frequency of choosing the unfamiliar object depending on the condition. The bold horizontal lines represent the mean, while the rectangular boxes denote the 95% highest density interval (HDI) around the mean. The width of the beans is in an indicator of the density with the dots representing individual data points. The upper right panel shows the prior and posterior distribution of the effect size *d* along with the posterior 95% HDI. The lower panel shows the results of the robustness analysis. The Bayes Factor (BF) is plotted as a function of the scaling factor of the prior on the effect size. Dashed gray lines mark different conventional evidence thresholds for accepting H_1 and H_0 while the bold gray lines represent the evidence thresholds of 5 and 1/5 we chose for our analysis

mappings when pairs of objects were introduced with two novel labels in each trial, so long as the word-object mappings were maintained consistent across trials. In contrast, we found that the initial BF crossed the threshold for the null hypothesis at 32 participants (compared to 55 participants in the original study). At this point, we could have stopped testing since we had evidence that the data was five times more likely under the hypothesis that the effect was exactly zero than under the alternate hypothesis of a non-zero effect. Consider also that our BFDA showed very low false negative (i.e., type II error) rates even for relatively small effect sizes as well as acceptable true negatives rates in case of a true effect size of zero (see Appendix A). Nevertheless, we decided to continue testing further given that our results were contrary to previous studies finding evidence for H_1 in such paradigms. At 43 participants, we stopped testing with the BF supporting the null hypothesis. As opposed to frequentist analyses, the Bayesian analyses reported here quantified evidence for the null hypothesis. In other words, there was no effect of learning of the word-object associations. Quantifying evidence for the null is a known benefit of Bayesian analyses, but more so in SBF designs because of the flexibility in continuing testing until such evidence is obtained is a clear plus, in our opinion. We will not go into details with regards to potential reasons for not replicating this finding, but we suspect that the age of the participants may be crucial here since most replication attempts have tested older children (see Appendix B).

There were a number of advantages to using an SBF design with regards to this replication attempt. First, after testing our initial set of 20 participants, SBF allowed us to continue testing until we crossed the lower threshold rather than reporting an inconclusive result. Thus, we were able to use the data collected from the first 20 participants towards a final result that could be interpreted, rather than ending up with an inconclusive result. In our opinion, this will lead to fewer datasets being relegated to the file-drawer if researchers are able to collect more data with the hope of a more interpretable result. Second, we allowed ourselves further flexibility in continuing testing even after we crossed the threshold, because it was important to us to ensure that the BF truly stayed past the threshold and that the evidence in favor of the null hypothesis was not a stray negative result (rare as they might be in the case of the design we chose). This is the true benefit of this design, because the BF under the null hypothesis meanders to zero with increasing sample size (*consistency*, Morey & Rouder, 2011; Rouder et al., 2012). Therefore, testing more participants would only provide stronger evidence for the null hypothesis, were this to be true.

Finally, we briefly examine our replication of the Markman and Wachtel (1988) study. First, we note that we found strong evidence (BF >10) in favor of the alternative hypothesis, that children chose the unfamiliar object more often when presented with an unfamiliar label relative to being asked to select one of the objects. However, we noticed as we wrote up the results that we had made an error in planning the study and only tested 20

participants overall (i.e., 10 per condition, same *n* as the original study), rather than testing 20 participants per cell as recommended by Schönbrodt et al. (2017; although we note that we tested as many participants as the original study). We chose to report the results of the study for several reasons. First, despite the small sample, the study contributes to the strength of evidence for the mutual exclusivity bias. Were it to be included in a meta-analysis on the mutual exclusivity effect, it would still contribute to the estimation of the effect size (Fan et al., 2004; Schönbrodt et al., 2017). Second, the impact of the starting sample size on the frequency of false positives or false negatives is surprisingly small (Schönbrodt et al., 2017). In our case, a simulation yielded expected false positive rates of about 4% for a starting sample size of 10 per cell compared to a false positive rate of 3% when starting with a sample size of 20 per cell.[8] Thus, we report these results with the caveat that the smaller initial sample size may have inflated the rate of false positive evidence.

We do, however, note difficulties with regards to planning testing sessions when participants are booked in advance. This might entail that additional children are booked in to participate in a study, despite the fact that the BF has already crossed the threshold. On the one hand, it would be ethical to cancel the visit of the planned participant since their resources are strictly not required for the study. On the other hand, caregivers may become frustrated with testing sessions that are repeatedly cancelled. An alternative approach would be to only book in participants when certain that more are required, but this would lead to unnecessary delays in testing. A further solution may be to choose an incremental sample size >1 (or 1 per cell), so that the study still reaps the benefit of efficiency but does not need to send too many participants home without testing.

Finally, we note that an SBF design may also yield an inconclusive result if other factors (personnel, funding, time) entail a maximum sample size. For instance, in Study 2, we ceased testing at 43 participants because the researcher in-charge of testing was no longer available. Had the BF in this study still been between the thresholds for the null and the alternative hypothesis (i.e., BF between 1/5 and 5) at this point, we would have been left with an inconclusive result. This is similar to the problems faced by fixed-*n* designs, with the difference being that the restriction on sample size is due to resource limitations. Nevertheless, we suggest that an inconclusive result in an SBF design is considerably more useful than inconclusive frequentist results when considered from the perspective of cumulative science. Specifically, we could have trained another experimenter and continued data collection at a later point. Alternatively, one could use the posterior of such an inconclusive result and use it to model an informed prior for a follow-up study.

In conclusion, in the current study, we replicated using an SBF design two previously reported effects, namely the mispronunciation effect (Swingley & Aslin, 2000) and the mutual exclusivity effect (Markman & Wachtel, 1988). We did not replicate the cross-situational learning effect in our sample of German 12- to 14-month-olds (Smith & Yu, 2008). Across these replication attempts, we see clear benefits of using an SBF design in developmental research while acknowledging some minor issues that are yet to be resolved. The final results we report here upheld the criteria of informativeness and efficiency (Stefan et al., 2019). As we have argued above, informativeness and efficiency may be particularly important to developmental research, given the demands on already overworked caregivers to participate in our research as well as the time and effort required on the part of researchers. The efficiency of SBF designs—Schönbrodt et al. (2017) suggests such designs may be between 50% to 70% more efficient—may be particularly important to developmental research which is, as noted above, plagued by issues of smaller sample sizes and may be especially attractive to researchers with limited funding.

ACKNOWLEDGEMENTS

This work was funded by a seed grant awarded to NM and TS by the Leibniz Science Campus Primate Cognition.

CONFLICTS OF INTEREST

There are no conflicts of interest to declare.

DATA AVAILABILITY STATEMENT

The data supporting the findings of this study are openly available in OSF https://osf.io/kpsy3/.

ORCID

Melanie S. Schreiner https://orcid.org/0000-0002-1530-8839

ENDNOTES

[1] Researchers who consider one hypothesis more likely, a priori, can easily adapt the methods described here. They can simply determine the BF required to reach the thresholds for accepting H_1 and H_0, respectively, using the formula shown above.

[2] This dissociation may be better understood against the context of how inference criteria in frequentist and Bayesian statistics develop over time as a function of increasing sample size. If the alternative hypothesis is true, the Bayes Factor typically converges to infinity with increasing sample size. On the other hand, if the null hypothesis is true, the Bayes Factor converges to zero with increasing sample size. Thus, increasing sample size can only strengthen evidence for the true hypothesis in Bayesian statistics, a property referred to as consistency (Morey & Rouder, 2011; Rouder et al., 2012). In contrast, in frequentist analyses, while the *p* value does converge to 0 were the alternative hypothesis to be true, it does not follow a systematic pattern were the null hypothesis to be true. In the latter case, the *p* value takes a random value between 0 and 1, such that were an infinite value of tests conducted, some of them would inevitably lead to false significant results.

[3] Note that Gronau et al. (2020) suggest a relatively simple approach to define subjective priors that allows quantifying the departure from predictive matching and information consistency.

[4] For consistency, we also report here the analyses with the reduced set of 20 children for consistency, $t(19) = 3.49$, BF = 17.01, $d = 0.78$, 95% HDI [0.22, 1.21].

[5] A reviewer pointed out that it would have been justified to use one-tailed tests when replicating previously published effects. We re-ran

the analysis using a one-tailed Bayesian *t*-test (the only difference to a two-tailed tests is that that the prior for the effect size under H_1 is a so called half-Cauchy, that is, only positive values are allowed). Using the one-tailed test yielded evidence consistent with the two-tailed version but the evidence for H_1 was substantially stronger, $t(20) = 3.10$, BF = 16.19, $d = 0.61$, 95% HDI [0.15, 1.06].

[6] We again ran a one-tailed version of our test. Had we used one-tailed *t*-tests, we would have stopped collecting data at a sample size of 33 participants, $t(32) = 0.06$, BF = 0.195, $d = 0.14$, 95% HDI [0.00, 0.33]. Note that the estimate of the effect size seems larger than in the two-tailed test because possible values for the effect size are restricted to be positive in the one-tailed test. For our final sample size, the one-tailed test would have been inconclusive, $t(42) = 0.43$, BF = 0.24, $d = 0.14$, 95% HDI [0.00, 0.33].

[7] A one-tailed Bayesian *t*-test would have yielded even stronger support for H_1, $t(18) = 4.55$, BF = 186.56, $d = 1.76$, 95% HDI [0.63; 2.91].

[8] In this simulation, we used default Bayesian *t*-tests for independent samples. We generated data from normal distributions, assuming that H_0 was true. Starting sample size was either 10 or 20 per cell, and this sample size was increased in increments of 5 up to a maximum of 50 per cell or until the BF exceeded 5 or fell below 1/5.

REFERENCES

Ackermann, L., Hepach, R., & Mani, N. (2020). Children learn words easier when they are interested in the category to which the word belongs. *Developmental Science*, *23*(3), e12915. https://doi.org/10.1111/desc.12915

Armitage, P., McPherson, C. K., & Rowe, B. C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society. Series A (General)*, *132*, 235–244.

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

Fan, X., DeMets, D. L., & Lan, K. K. G. (2004). Conditional bias of point estimates following a group sequential test. *Journal of Biopharmaceutical Statistics*, *14*, 505–530.

Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (2020). Informed Bayesian t-tests. *The American Statistician*, *74*, 137–143.

Jeffreys, H. (1961). *The theory of probability*. Oxford University Press.

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meaning of words. *Cognitive Psychology*, *20*(2), 121–157.

Miller, S. A. (2012). *Developmental research methods*. Sage.

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*, 406–419.

Oakes, L. M. (2017). Sample size, statistical power, and false conclusions in infant looking-time research. *Infancy*, *22*, 436–469.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*, 356–374.

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

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*, 128–142.

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*, 322–339.

Smith, L. B., & Yu, C. (2008). Infants rapidly learn word- referent mappings via cross-situational statistics. *Cognition*, *106*, 1558–1568.

Stefan, A. M., Gronau, Q. F., Schönbrodt, F. D., & Wagenmakers, E. J. (2019). A tutorial on Bayes Factor Design Analysis with informed priors. *Behavior Research Methods*, *51*, 1042–1058.

Swingley, D., & Aslin, R. N. (2000). Spoken word recognition and lexical representation in very young children. *Cognition*, *76*(2), 147–166.

Swingley, D., Pinto, J. P., & Fernald, A. (1999). Continuous processing in word recognition at 24 months. *Cognition*, *71*, 73–108.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E. J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives in Psychological Science*, *6*(3), 291–298.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Mani N, Schreiner MS, Brase J, et al. Sequential Bayes Factor designs in developmental research: Studies on early word learning. *Dev Sci.* 2021;24:e13097. https://doi.org/10.1111/desc.13097