languages

MDPI

Article

Even Simultaneous Bilinguals Do Not Reach Monolingual Levels of Proficiency in Syntax

Wei Li and Joshua K. Hartshorne *

Department of Psychology and Neuroscience, Boston College, 522 McGuinn Hall, Chestnut Hill, MA 02467, USA
* Correspondence: joshua.hartshorne@bc.edu

Abstract: While there is no doubt that children raised bilingual can become extremely proficient in both languages, theorists are divided on whether bilingualism is effectively monolingualism twice (the "Two Monolinguals in One Brain" hypothesis) or differs in some fundamental way from monolingualism. A strong version of the "Two Monolinguals" hypothesis predicts that bilinguals can achieve monolingual-level proficiency in either (or both) of their languages. Recently, Bylund and Abrahamsson argued that evidence of lower syntactic proficiency in simultaneous bilinguals was due to confounds of language dominance; when simultaneous bilinguals are tested in their primary language, any difference disappears. We find no evidence for this hypothesis. Meta-analysis and Monte Carlo simulation show that variation in published results is fully consistent with sampling error, with no evidence that method mattered. Meta-analytic estimates strongly indicate lower syntactic performance by simultaneous bilinguals relative to monolinguals. Re-analysis of a large dataset (N = 115,020) confirms this finding, even controlling for language dominance. Interestingly, the effect is relatively small, challenging current theories.

Keywords: second language acquisition; cricial period; bilingualism

check for updates

Citation: Li, Wei, and Joshua K. Hartshorne. 2022. Even Simultaneous Bilinguals Do Not Reach Monolingual Levels of Proficiency in Syntax. *Languages* 7: 293. https://doi.org/10.3390/languages7040293

Academic Editors: David P Birdsong, Juana M. Liceras and Raquel Fernández Fuertes

Received: 4 June 2021
Accepted: 17 March 2022
Published: 16 November 2022

1. Introduction

Perhaps the only thing more remarkable than the fact that humans can learn a language—something that no other animal or machine can manage—is that we can learn more than one. It is not pre-theoretically obvious that just because we can learn one language, we should also be able to learn two.

The problem is not just fitting two languages into one brain, though how the brain represents two separate communication systems is something that needs explanation. Being a bilingual is different from being a monolingual and presents additional challenges. Consider the scenarios that bilinguals encounter many times every day: listening to people talk, speaking with others, and reading written materials. The bilingual has to identify which language is being used in the scenario and then choose the appropriate lexicon as well as the corresponding syntax—all in fractions of a second.

Moreover, one might naively assume that it would take twice as long to learn two languages as one, or that bilinguals would learn each of their languages to only 50% proficiency. This is clearly not the case: given the right environment (namely, regular use for each language), bilinguals can achieve very high levels of proficiency in both languages, and in nearly the same time frame as that of monolinguals (for obvious reasons, bilinguals who rarely need to use one of their two languages tend to suffer in proficiency in that language). This is only more impressive for individuals who know more than two languages. (Note that because the present paper is about the difficulties of knowing more than one language, our primary point of contrast is between people who know one language and people who know more than one. To avoid unnecessarily verbose writing, we will speak of "bilinguals" who know "two languages" rather than "bilinguals or multilinguals" and "two or more". Wherever the differences between knowing two languages and knowing more is critical, we will be explicit in our terminology.)

What remains unclear is whether bilinguals can achieve *the same* level of proficiency as monolinguals, or merely very similar. The core theoretical implication here is whether bilingual acquisition can be understood as monolingual acquisition twice (the Two Monolinguals in One Brain Hypothesis), or whether bilingualism is its own (at least subtly) different phenomenon (Birdsong and Gertken 2013; Grosjean 1989). If the latter is true, it would raise further theoretical questions (*What is the nature of this difference? What are its causes?*) which we return to in the General Discussion.

There are practical implications as well. Studies of second language learners often compare them to monolinguals. This confounds age of acquisition with bilingualism, potentially judging later learners against an unfair standard and artificially inflating differences between early and late learners (Birdsong 2005). Similarly, in clinical settings, bilinguals are often compared against monolingual norms (Lakshmanan 2013; Thordardottir et al. 2006). Again, if the average bilingual does not reach the average monolingual norm, this method risks spurious identification of language deficits.

1.1. Two Monolinguals in One Brain Hypothesis: State of the (Syntactic) Evidence

In the present paper, we focus on syntax. Syntax is particularly interesting because of the clear theoretical plausibility of no "bilingual difference" for syntax. Syntax is often argued to be mastered by children within a few years—particularly under Nativist theories (Crain 1991; Pinker 1994; Wexler 1998)—and thus simultaneous bilinguals (those who learn two languages from birth, also called "crib bilinguals") should still have plenty of time to obtain enough input in both languages.

Contrast this with vocabulary, where there is effectively an unbounded number of words one could learn, most of which are exceedingly rare (Zipf 1935). A great deal of input is required just to plausibly encounter low-frequency words, and indeed, even monolinguals continue acquiring new words into their seventh decade of life (Hartshorne and Germine 2015). Thus it is probably unreasonable to expect bilinguals to have the same size vocabulary in a given language as a monolingual does, simply due to lessened opportunity.

Studies to date have differed on whether simultaneous bilinguals can reach monolingual levels of syntactic knowledge (Ardila et al. 2019; Bylund et al. 2020; Garraffa et al. 2017, 2020) or not (Giguere and Hoff 2020; Hartshorne et al. 2018). Some authors have suggested that the differences across studies can be explained by differences in methodology. In particular, Bylund et al. (2020) recently argued that the findings of a "bilingual decrement" (sub-monolingual performance on the part of simultaneous bilinguals) are due to a confound, at least in the case of syntax. Specifically, they argue that where a bilingual decrement for syntax has been observed, it is due to including bilinguals for whom the target language was not their primary language. They report a study of heritage Spanish speakers in Sweden, reporting that their performance on a Swedish grammaticality judgment task was not significantly different from that of monolingual Swedish speakers. (Throughout, we use the term "heritage bilinguals" to mean a bilinguals who learned a "foreign" language from their parent(s), as well as the local majority language.)

A more parsimonious explanation is statistical power. The highest-powered study to date, from Hartshorne et al. (2018), reported that bilinguals from birth scored 0.17 standard deviations below monolinguals. The sample Bylund and colleagues used (20 monolingual and 20 bilinguals) was only large enough to detect an effect of that magnitude 7.55% of the time (as estimated in a standard power analysis for *t*-tests). That is to say, we would expect Bylund and colleagues to report a null result *whether or not* their hypothesis is true. To quantify this, if we thought *a priori* that Bylund and colleagues had a 50% chance of being correct, then observing a null result should only increase our confidence to 50.6% [1].

This uncertainty can also be seen in the inspection of confidence intervals. Though not highlighted by the authors, Bylund and colleagues' estimate of how much more poorly simultaneous bilinguals perform on a grammaticality judgement task runs from 0.99 standard deviations (larger than the vast majority of effects in psychology; Hartshorne and Schachner (2012); Stanley et al. (2018)) to a bilingual *advantage* of 0.26 standard deviations. Indeed, despite interpreting their results as reflecting a lack of difference between monolinguals and simultaneous bilinguals, their point estimate of the "bilingual decrement" (Cohen's $d = 0.36$) is one of the largest in the literature, nearly twice what was reported by Hartshorne and colleagues.

1.2. Overview of the Present Paper

The goal of the present study is to resolve the dispute in the literature as to whether simultaneous bilinguals reach monolingual levels of syntactic knowledge. We take three complementary approaches.

Studies 1 and 2 estimate how much of the variability in the literature can be explained purely by statistical noise. In Study 1, we report a meta-analysis of prior work, evaluating evidence that discrepancies in results are beyond what could be explained by statistical chance (for review and discussion of meta-analysis, see Campbell et al. 2020; also Plonsky et al. 2021). Meta-analysis takes advantage of all the available data, but is limited by any publication biases, the assumptions of parametric statistics, and by how many studies have been conducted (in this case, not many).

In Study 2, we report a Monte Carlo simulation of one prior study, again in order to estimate just how much variability across studies we should expect from statistical noise alone. While Monte Carlo simulation depends on only a subset of the available data, it has the advantage in that it is nonparametric and so depends on fewer statistical assumptions. Converging results across Studies 1 and 2 would provide greater confidence in each.

Finally, we directly test the hypothesis raised by Bylund and colleagues: that when simultaneous bilinguals are tested in their primary language—and when that primary language is also the majority community language—they are indistinguishable from monolinguals. We conduct a high-powered test of this hypothesis by re-analyzing data from Hartshorne et al. (2018).

2. Study 1: Meta-Analysis

2.1. Target Papers

Our literature search was consistent with the guidelines advocated by PRISMA 2020 (Page et al. 2021). Because our interest is in syntactic proficiency averaged across the whole of syntax, we restricted the meta-analysis to papers with a broad measure of syntactic proficiency or which measured a large number of different syntactic phenomena (e.g., passivization, clefting, agreement, relative clauses, preposition use, verb syntactic subcategorization, pronoun gender and case, modals, determiners, subject-dropping, aspect, sequence of tenses, and wh-movement). We also restricted the search to studies of adult bilinguals, because the question is not whether bilinguals learn both of their languages as quickly as monolinguals learn their one, but whether they reach the same level of proficiency. We counted a study as including simultaneous bilinguals if the study labeled the group as such or had a group described as being exposed to two languages from birth. Methods could involve judgment, acceptability, or picture matching. The dependent variables could be accuracy, reaction time, Likert scales, or similar.

During the search, we were unable to identify keywords that were effective at identifying papers of interest. Thus, in addition to a number of different search terms (for which we considered anywhere from a couple dozen to a couple hundred results), we also adopted a recursive strategy: for every paper added to our set, we investigated everything it cited or that cited it. We also solicited suggestions and unpublished data from professional listservs, and reached out directly to a number of experts in the field. We were able to identify 10 experiments from 6 papers, all of which were published in journals.

2.2. Data Entry

Data were entered by the first author. Two experiments used a Likert scale, two experiments reported reaction times, and six reported accuracy. In each case, we were able to extract a mean and variance separately for monolinguals and simultaneous bilinguals, calculated using the same units used in the original paper. Effect sizes were then calculated in terms of Cohen's d, using a pooled standard deviation (Cohen 1988).

An argument could be made for log-transforming the reaction times and converting accuracies to log-odds (only one study reported accuracy in log-odds) (cf. Jaeger 2008). However, such transformations have their greatest benefit when conducted prior to averaging across subjects, and they matter most for interactions, which, with one minor exception, we are not testing. Thus, for simplicity, we did not do any more transformations and just used the original units that were used in each study.

2.3. Results

All the analyses were conducted in R. We conducted a random effects meta-analysis (Balduzzi et al. 2019; Fleiss 1993). The Forest plot is shown in Figure 1. The standardized mean difference in monolinguals and simultaneous bilinguals and 95% confidence interval whiskers are indicated for each study. As can be seen, the confidence intervals are largely overlapping and in most cases include the estimated means for the other studies—the primary exception being Hartshorne et al. (2018), which has a tiny confidence interval as a result of its large sample. Figure 1 also presents the 95% prediction interval, within which the effect sizes of 95% of future studies should fall, assuming there is no selection bias. It indicates that, given typical sample sizes, both studies showing a bilingual decrement and studies showing no bilingual decrement are expected.

The meta-analytic estimated effect is 0.19, right at the threshold of significance ($p = 0.09$; 95% CI = (-0.04, 0.42)). That is, in aggregate, the published literature suggests a small but non-trivial bilingual decrement. Although Hartshorne et al. (2018), by virtue of containing almost all of the data in the literature, has an outsized influence on the meta-analysis, excluding it did not meaningfully change the results beyond increasing uncertainty.[2]

Thus, statistical analysis suggests that that the differences in results across studies can be explained largely by statistical noise. Heterogeneity analysis suggested a great degree of similarity between the studies and that methodological differences had little effect on results ($I^2 = 0.37$, $p = 0.11$). The Funnel plot, Figure 2, shows the classic funnel shape, with the study with the most power and precision sitting right at the mean of the distribution.

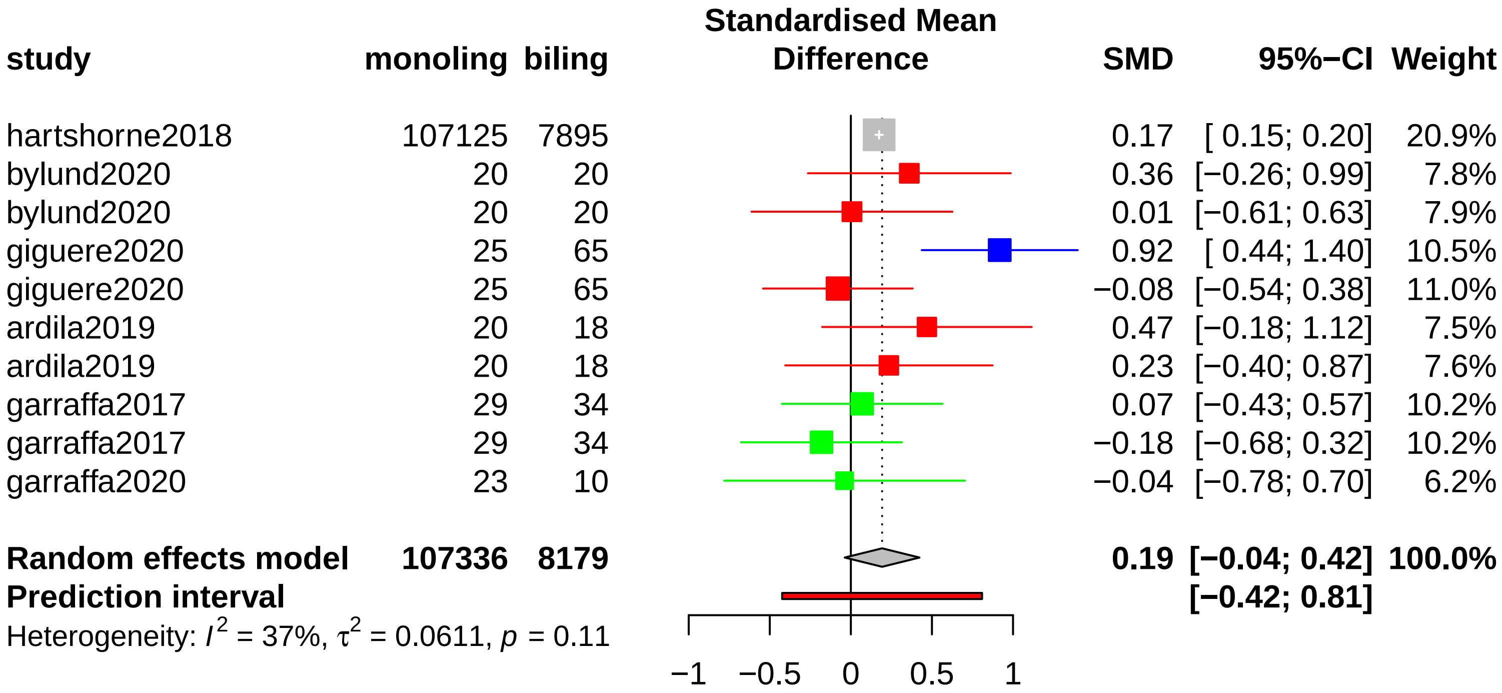

Figure 1. Meta-analysis forest plot. *Red*: heritage bilingual (at least one parent speaks 'foreign' language), tested in majority language. *Blue*: heritage bilingual, tested in heritage language. *Green*: minority language (member of non-immigrant minority language group), tested in majority language. *Grey*: mixed (subjects involve some mix of the other categories) (Ardila et al. 2019; Bylund et al. 2020; Giguere and Hoff 2020; Hartshorne et al. 2018; Garraffa et al. 2017, 2020).

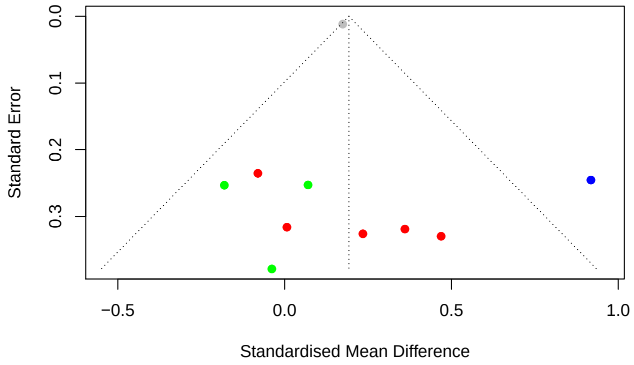

Figure 2. Meta-analysis funnel plot. Red: heritage bilingual (at least one parent speaks 'foreign' language), tested in majority language. Blue: heritage bilingual, tested in heritage language. Green: minority language (member of non-immigrant minority language group), tested in majority language. Grey: mixed (subjects involve some mix of the other categories).

2.4. Discussion

Statistical analysis of the literature revealed that although different studies have reported different levels of significance, there is little to no evidence of a statistically significant difference *between* studies.

However, this is more an absence of evidence than evidence of absence, since there are methodological differences within the dataset that probably do matter. For instance, the two datapoints from Garraffa et al. (2017) involve bilingual speakers of languages that are so closely related (Italian and Sardinian), one might reasonably expect minimal effects of bilingualism. Indeed, these datapoints are close to 0. However, they are not statistically different from the rest of the dataset.

On the other end of the spectrum, Giguere and Hoff (2020) is the only study to report data from the non-majority language. In particular, this analysis involved Spanish–English American bilinguals who had not been educated in Spanish. One might reasonably expect their Spanish proficiency to be less than that of monolinguals, and indeed they show the largest bilingual decrement. However, it is at best marginally different from the rest of the studies, with its point estimate within the confidence intervals of several of the other studies—including, notably, Bylund et al. (2020). Moreover, statistical analysis does not indicate that these results constitute a significant outlier.

Thus, the clearest conclusion is that any effects of the methodological differences between studies in the literature are too small to be detectable, given the fairly large error bars of these studies. Interestingly, the only study that has small error bars—Hartshorne et al. (2018)—also happens to provide an effect size estimate right at the meta-analytic mean of the other studies. Thus, the literature as it stands provides weak evidence for a bilingual decrement.

3. Study 2: Monte Carlo Simulations

3.1. Bootstrapping from the Data of Hartshorne and Colleagues

Another method of investigating whether the results of prior studies differed more than would be expected by chance is Monte Carlo simulation. Specifically, we used the data from Hartshorne et al. (2018) (publicly available at http://osf.io/pyb8s (accessed on 23 February 2022)) to simulate re-running the English data from Giguere and Hoff (2020). We chose this study for three reasons. First, it is the largest study other than Hartshorne et al. (2018) itself; it therefore represents a relatively "high-powered" study by the standards of the literature. Second, unlike either Hartshorne et al. (2018) or Bylund et al. (2020), its grammaticality judgment task involves a Likert scale rather than binary judgment (cf. Langsford et al. 2018). Including it therefore allows us to get a sense of how effect sizes transfer across the exact method used. Third and finally, it did not show a significant bilingual decrement, as predicted by Bylund et al. (2020).

3.2. Method

Hartshorne et al. (2018) report the results of a 10-minute online English grammar quiz, consisting of 132 items covering a broad range of syntactic phenomena, including passivization, clefting, agreement, relative clauses, preposition use, verb syntactic subcategorization, pronoun gender and case, modals, determiners, subject-dropping, aspect, sequence of tenses, and *wh*-movement, among others. There were 124 two-alternative forced-choice (2AFC) decisions and 2 four-alternative forced-choice (4AFC) decisions, though for analytic reasons the latter were scored as 8 2AFC decisions, for a total of 132. Of these, 124 involved grammaticality judgment and 8 involved sentence–picture matching. In order to improve the subject experience, the grammaticality judgment questions were presented in sets of 4 or 8. The data were then used to guess the subject's native language and dialect of English. Subsequent to viewing the guess, subjects were invited to provide demographic information in order to help improve the model, including their native language, primary language, years of schooling in English, and other key demographic factors. Participants found the quiz very engaging and it was virally shared on social media. Hartshorne et al. (2018) report data from the 669,498 participants who completed the study, excluding repeats and implausible responses. Note that in Hartshorne et al. (2018), all analyses are conducted in terms of log-odds instead of percent correct, because the latter artificially imposes ceiling effects. Additional details about the method, including a full list of stimuli, are included in Hartshorne et al. (2018).

We simulated re-running the English-language study in Giguere and Hoff (2020), following the reported demographics as closely as possible. They tested 25 English monolinguals and 65 English–Spanish simultaneous bilinguals. Most subjects were college students, with average ages of 19.1 and and 19.5, respectively. Thus, we restricted our data to subjects aged 18 to 25 who reported their maximum level of education as being "some college" (we do not know whether the individuals were currently enrolled, but given the age restriction, this is probably close enough). Following Giguere and Hoff (2020), we restricted the data to subjects born in the United States and currently living in the United States. Further following Giguere and Hoff (2020), we restricted monolingual English speakers to individuals who did not learn any language other than English from birth and whose current primary language is English. Giguere and Hoff (2020) further excludes monolinguals with 5% or more exposure to any non-English language, but since Hartshorne et al. (2018) did not record such data we cannot use them for exclusion.

For the Spanish–English bilinguals, we followed Giguere and Hoff (2020) in restricting the bilingual sample such that 1 had never attended university, 4 had graduated university, the remaining 60 were college students (again defined for our purposes as listing "some college" as their maximal level of education). Giguere and Hoff (2020) defines Spanish–English bilinguals as having begun learning both languages prior to the age of 5. While Hartshorne et al. (2018) records the exact age of English acquisition, acquisition of non-English languages is dichotomized as being learned "from birth" or not. Thus, we used a slightly stronger criterion: learning Spanish from birth and English before the age of 5.

In practice, this likely makes no difference: Giguere and Hoff (2020) recruited Spanish-speakers born in the United States; it is unlikely that many of them only began learning Spanish after starting English. However, we ran a subsequent analysis where subjects were required to have learned both languages from birth.

The resulting sample included 13994 monolinguals, 817 Spanish–English bilinguals for the first simulation (simultaneous bilinguals with AoA smaller than 5), and 717 for the second simulation with stronger criteria (simultaneous bilinguals with AoA is 0). We conducted 1000 simulated experiments. For each experiment, we drew 25 monolingual and 65 bilinguals subjects with replacement from the restricted data set described above. We then estimated the size of the bilingual decrement in terms of Cohen's d, again using pooled variance.

3.3. Results and Discussion

The distribution of the effect size of the simulated experiments can be seen in Figure 3. In the primary simulation, most of the prior studies fell within the 95% confidence interval (−0.06, 0.75; Figure 3, top), including studies that showed a significant bilingual decrement and those that did not. Results were similar for our follow-up analysis, which used the stronger criterion for simultaneous bilinguals of both languages who reported learning from birth (−0.06, 0.72; Figure 3, bottom).

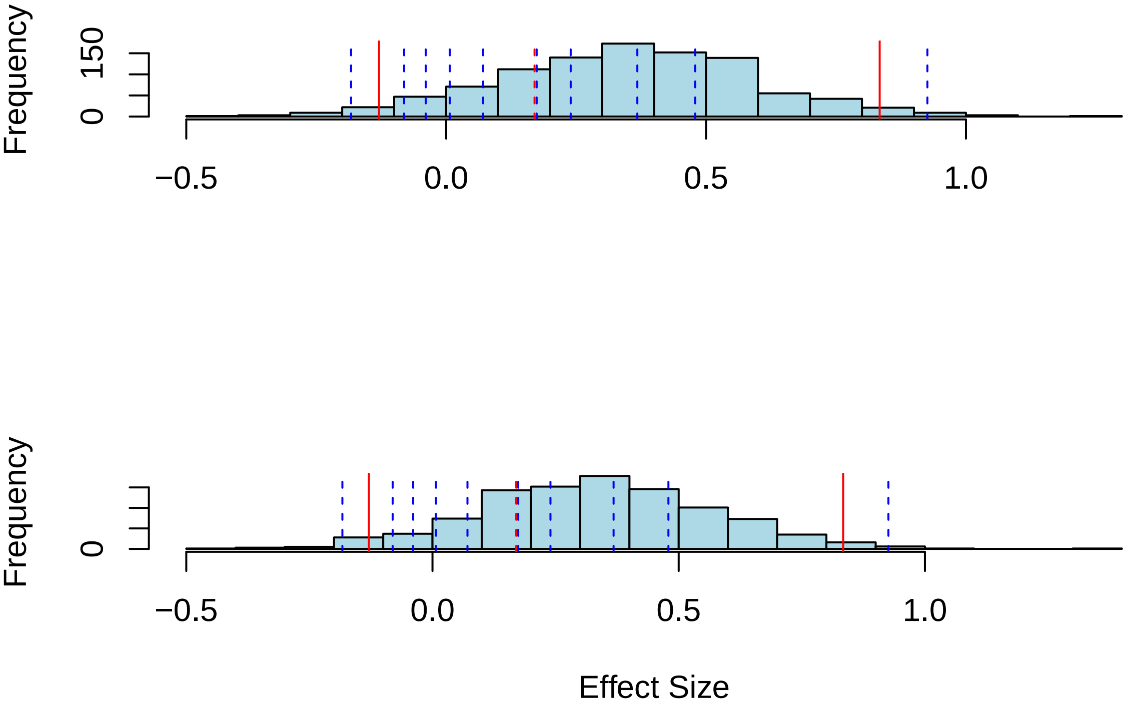

Figure 3. Histograms of simulation results. X-axis is size of bilingual decrement in terms of Cohen's d (negative numbers indicate a numerical advantage for bilinguals). Solid red lines show the 95% confidence interval. Dashed lines show the measured effect sizes from studies included in the meta-analysis (the red dashed line is Hartshorne et al. 2018). Top: Using Giguere and Hoff's (2020) criterion such that subjects began learning both languages before the age of 5. Bottom: Using the stronger criterion that both languages were learned from birth.

Critically, the Giguere and Hoff (2020) English data were themselves within the 95% confidence interval, despite the fact that Giguere and Hoff (2020) used a different response measure (Likert) (for discussion, see also Langsford et al. 2018). This suggests that—with the exception of Hartshorne et al. (2018)—differences in the results of prior studies are well within what one expects due to random statistical noise. This is consistent with the findings of the meta-analysis.

Interestingly, the best estimate of the effect size from both simulations was substantially larger than what was reported by Hartshorne et al. (2018) based on their entire dataset. Further investigation showed that this was because Spanish–English bilinguals show a larger bilingual decrement than the average simultaneous bilingual in Hartshorne et al. (2018).

4. Study 3: Comparing Simultaneous Bilinguals Whose Primary Language Is or Is Not English

Bylund et al. (2020) suggest that Hartshorne et al. (2018) observed a bilingual decrement due to inclusion of simultaneous bilinguals whose current primary language was not English and who did not grow up in an English-speaking context. We tested this by reanalyzing the data from Hartshorne et al. (2018).

4.1. Method

We used the same blanket exclusions used by Hartshorne et al. (2018). We then further restricted the sample to those who reported that currently their sole primary language is English, and that they had only ever lived in English-speaking countries (defined as Australia, Canada, United Kingdom, United States, New Zealand, Ireland (Republic of), Singapore, South Africa, India). From the remaining subjects, we identified monolingual English speakers (those who reported learning only English from birth) and simultaneous bilinguals (those who learned two or more languages from birth, one of which must have been English).[3]

In principle, we could compare the entire lifespan trajectories for the two learner groups. However, comparing lifespan trajectories is a notoriously thorny and largely unsolved statistical problem (Hartshorne and Germine 2015). Thus we instead focused on asymptotic performance: the level of performance typical of a mature speaker who is no longer improving measurably in their use of syntax. Thus, we excluded subjects under the age of 30, the approximate age at which participants reach asymptote in this quiz (Hartshorne et al. 2018). We further excluded subjects older than 70, who may be showing evidence of cognitive decline (Hartshorne et al. 2018).

The final dataset included 73163 monolinguals (age = 42.54, sd = 10.41), and 2981 English-primary simultaneous bilinguals (age = 40.32, sd = 9.66).

4.2. Results and Discussion

Results are shown in Figure 4. Performance on the syntax test was higher for monolinguals (M = 3.60, SE = 0.00) than for simultaneous bilinguals whose primary language was English and who have always lived in an English-speaking country (M = 3.48, SE = 0.02), a result that reached significance in a two-sample *t*-test ($t(3, 230.50) = 8.19$, $p < 0.001$, Cohen's *d* = 0.15). This effect size was only slightly smaller than what was reported in Hartshorne et al. (2018) using all subjects (Cohen's *d* = 0.17). Thus, we find no evidence for the hypothesis raised by Bylund et al. (2020): simultaneous bilinguals are indistinguishable from monolinguals, provided that they are tested in their primary language and that this language is the primary language in their community from birth. [4]

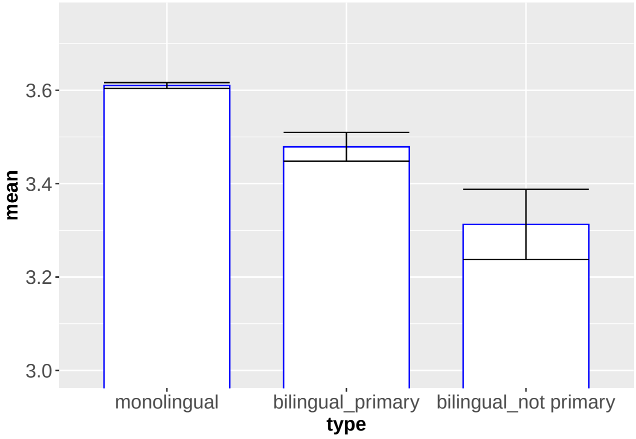

Figure 4. Re-analysis of Hartshorne et al. (2018), comparing simultaneous bilinguals whose primary language is or is not English.

We ran a follow-up analysis to test whether being a heritage speaker or having English as a primary language actually affects English proficiency for simultaneous bilinguals. For this analysis, we considered simultaneous bilinguals ages 30 to 70, regardless of their primary language, and we restricted them to those who had *only* lived in English-speaking countries or who had *never* lived in English-speaking countries. (It might have been informative to consider proportion of time spent in English-speaking countries, but Hartshorne and colleagues did not systematically collect this information for native speakers of English.) We subjected the resulting 4958 simultaneous bilinguals to a linear regression with the two predictors (primary language and heritage status) and their interaction. Having English as one's sole primary language resulted in a substantial increase in English syntactic knowledge, as expected ($b = 0.19$, 95% CI $(0.09, 0.30)$, $t(4954) = 3.59$, $p < 0.001$, Cohen's $d = 0.10$, CI = (0.05, 0.16)). Living only in an English-speaking country (as opposed to never) only marginally improved syntactic knowledge ($b = 0.08$, 95% CI $(0.00, 0.17)$, $t(4954) = 1.85$, p= 0.064, Cohen's $d = 0.05$, CI = (0.00, 0.11)). The interaction was not significant ($b = -0.09$, 95% CI $(-0.21, 0.03)$, $t(4954) = -1.46$, p= 0.145).

Thus, although the factors highlighted by Bylund and colleagues do affect English proficiency, their effects are quite small (at least in the data from Hartshorne and colleagues). Moreover, because the Hartshorne et al. (2018) data contained very few subjects who did not meet the more restrictive definition used by Bylund and colleagues, excluding them had only a minimal effect on the results.

5. General Discussion

Results to date have differed on whether simultaneous bilinguals can reach monolingual levels. Some studies have reported that simultaneous bilinguals could not reach monolingual levels (Giguere and Hoff 2020; Hartshorne et al. 2018) while some failed to find the difference (Ardila et al. 2019; Bylund et al. 2020; Garraffa et al. 2017, 2020). We tested whether this is best explained by chronically low statistical power or by methodological differences across the studies, with a focus on a suggestion by Bylund et al. (2020) that findings of a "bilingual decrement" (sub-monolingual performance on the part of simultaneous bilinguals) disappear when one considers simultaneous bilinguals learning in a heritage context tested in their dominant language, but only when their dominant language is also the majority language.

A meta-analysis (Study 1) showed that the differences in results across prior studies matched what would be expected due to random chance alone. In particular, the measured effect size for the largest study to date (Hartshorne et al. 2018) was a near perfect match for the meta-analytic combination of all the smaller studies—again, exactly what we would predict if statistical noise (but not methodological differences) explained variation in results. Further support comes from the fact that the parametric confidence intervals of the prior studies included the meta-analytic estimate in all but one case: the Spanish data from Giguere and Hoff (2020), which just barely missed the meta-analytic estimate. (In fact, in most cases, the CI for each study included the point estimate for every study in the meta-analysis.) That is, there is nothing for methodological differences across studies to explain.

Study 2 provided converging evidence for the hypothesis that the variable results across prior studies are due to statistical noise, using a nonparametric bootstrapping simulation. Specifically, we used the data from Hartshorne et al. (2018) to bootstrap confidence intervals for Giguere and Hoff (2020)—the second-largest prior study and the only one to (slightly) differ from the meta-analytic result in Study 1. The bootstrapped confidence interval was substantially wider than the parametric confidence interval and included nearly every study to date. Again, this suggests that (with the exception of Hartshorne et al. (2018)), the estimates provided by prior studies are so imprecise that there is no need to explain differences in results. Rather, they all produce the same, under-informative result.

Studies 1 and 2 cast doubt on the hypothesis of Bylund et al. (2020), but do not directly test it. Thus, Study 3 re-analyzed the data from Hartshorne et al. (2018) to directly test whether there is no difference in English syntactic knowledge between monolinguals and simultaneous bilinguals who grew up in an English-speaking country and who are dominant in English. In fact, there was a significant difference, nearly as large as the one initially reported by Hartshorne et al. (2018) (Cohen's $d = 0.15$ vs. 0.17). That is, even when simultaneous bilinguals are tested in their dominant language, that language happens to be the local majority language, and (by definition) they had learned that language from infancy, they still show a clear difference from monolinguals.

As noted in the introduction, the fact that bilinguals have difficulty achieving the same level of proficiency as monolinguals even in their dominant language is not pre-theoretically surprising: bilinguals have to learn two languages and navigate using two languages, and as a general rule, doing two things is harder than doing one. However, it is striking just how small the difference is. In some ways, this is the most confusing—and intriguing—outcome. The effect is too small to be easily explained by having two languages to learn (simultaneous bilinguals are nowhere near being only half as proficient as monolinguals), but it is also not negligible.

5.1. Interpreting Effect Size

An effect of 0.15 is on the small side for psychology, but well within the typical distribution (Gignac and Szodorai 2016; Hartshorne and Schachner 2012; Hemphill 2003; Richard et al. 2003; Stanley et al. 2018). Best estimates for the average effect size in psychology is around a Cohen's *d* of 0.4, with around a third of studies reporting effects of 0.2 or less.[5] However, these numbers are likely substantially inflated by the strong bias in psychology for publishing significant results (Bakker et al. 2012; Fanelli 2010; Sterling et al. 1995). Preregistered studies tend to report much smaller effect sizes, and replications—even when successful—do as well (Klein et al. 2019; Open Science Collaboration 2015; Schäfer and Schwarz 2019). Thus, the effect sizes we report here for the bilingual decrement (meta-analysis: 0.19; Hartshorne et al. (2018): 0.17; Study 3: 0.15) may not even be much below average.

To provide a benchmark, these effect sizes are similar to a difference of 2–3 IQ points—small, to be sure, but a pill that added two IQ points would sell quite well. For another benchmark, note that a widely-discussed effect in language acquisition and bilingualism is the effect of level of education. In the data from Hartshorne et al. (2018), the difference between mature monolinguals with some college education and those with a graduate degree and those with a high school education or less was Cohen's *d* = 0.27—less than twice what we observe for the bilingual decrement. (Comparing those with high school education and some graduate education, the effect is 0.38.)

It is important to keep in mind that standardized effect sizes like Cohen's *d* are calculated in terms of variation in the data. That is, the more noise in the data, the smaller the effect size. One reason effect sizes in psychology are small is that we typically accept a large amount of measurement error. That is certainly the case here: the short grammar tests used in the literature discussed in the present study (usually around 100 true/false questions) are coarse estimates at best. More precise instruments would necessarily lead to increased effect sizes.

5.2. Explaining High (But Below-Ceiling) Performance by Simultaneous Bilinguals

We already discussed why it would be easier to explain a large bilingual decrement than a small one. Similarly, explaining the lack of any bilingual decrement would be straightforward as well. For instance, theories where syntactic knowledge can be rapidly learned in the space of a few years would easily explain the lack of a difference, since simultaneous bilinguals have plenty of time to learn both (Crain 1991; Pinker 1994; Wexler 1998). However, positing that simultaneous bilinguals have almost-but-not-quite-enough time to learn both languages seems suspiciously exact.

One possibility is that simultaneous bilinguals do in fact have monolingual-level knowledge of at least one language, but there is slight interference between the languages resulting in subtle differences compared to monolingual norms (cf. Serratrice 2013). A slight variation of this account is that even well-honed skills become rusty if you do not use them for a while, and by definition bilinguals go a little longer between reusing any particular syntactic ability in a given language.

A second possibility derives from the Complementarity Principle (Grosjean 2016). The Complementarity Principle builds on the observation that bilinguals are not communicating with *either* of two languages; rather, they have a system to communicate with two languages. For some situations/interlocutors/topics/etc., bilinguals preferentially use one or the other language. Whereas monolinguals must use their own language for everything they do linguistically, bilinguals have a choice and may sometimes simply elect to not use a particular language for a particular use. This is most familiar for vocabulary: bilinguals may only know certain job-related words in the community language or certain religion-related words in a heritage language.

It is not implausible that complementarity might similarly apply to syntactic knowledge. For instance, in English, by-passives are a feature of academic writing and not colloquial speech (Street and Dąbrowsk 2010). Thus, a heritage speaker of English who has never used English in school may have had little cause to acquire by-passives. Perhaps our simultaneous bilinguals had (probabilistically) less experience with one relatively restricted syntactic phenomenon or another, and thus (as a group) performed a little below the monolingual norms. This is difficult to test with the data from Hartshorne et al. (2018): the wide coverage of syntactic phenomena means there is little data on any given syntactic phenomenon, making it difficult to distinguish differences between phenomena and simply random item effects (see: Hartshorne et al. 2018).

A third possibility is that it really does take a lot of input and practice to fully acquire the syntax of a language, and simultaneous bilinguals are woefully under-supplied with both. For instance, if there is a critical period for syntax acquisition (which is controversial), and if the critical period evolved in an environment where monolingualism was the norm (also controversial), there are good theoretical reasons to expect it to be only as long as is needed to fully acquire one language (Hurford 1991). However, bilinguals are able to learn with less data because of positive transfer between the languages (the existence of successful transfer across languages is well-documented, e.g., Snedeker et al. 2012).

The same principle may apply to syntax acquisition as well. There is growing evidence that there is influence on syntax across a simultaneous bilingual's languages (Serratrice 2013). Almost by necessity, this research has focused on cases where applying the patterns of one language to the other would result in a deviation from monolingual norms; it is less obvious how one would detect influence where both languages do the same thing. However, it stands to reason that if there is interference where the languages differ, there will also be mutual support where they are similar. If the support outweighs the interference, then a simultaneous bilingual will be able to learn two languages with less data than would be required by two monolinguals. The question is whether the support surplus would be sufficient to explain near-monolingual attainment.

6. Limitations and Conclusions

There are some limitations on the strength of our conclusions above, of which the most glaring is that most of the data involve the learning of English. While we have no particular reason to believe the story would be different for acquiring other languages, it would be far from the most surprising discovery to date. Another issue is that the bilingual decrement may be different—or even reversed—for different aspects of syntax, which is not something easily spotted or addressed in the data discussed above. Moreover, if this is the case, then measuring its overall effect requires testing a representative sample of syntactic phenomena. At the moment, constructing such a test is well beyond our theoretical understanding, much less our practical abilities. More broadly, we tested a specific hypothesis about the conditions under which simultaneous bilinguals might reliably reach monolingual levels. The fact that that hypothesis was disconfirmed does not rule out all other hypotheses.

As it stands, however, the analyses above add to the growing evidence that simultaneous bilinguals do not reach monolingual norms even in their primary language, but they do come exceedingly close. Both of these facts require explanation. Theoretical progress may come from a better empirical understanding of cross-language transfer, complementarity in syntax (if any), and the differences in input to monolingual and simultaneous bilingual learners.

Author Contributions: The authors made the following contributions. W.L.: data analysis, writing—original draft preparation, writing—review and editing; J.H.: conceptualization, writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This work was not supported by any funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data and analysis code can be found at https://osf.io/5smvd/, accessed on 23 February 2022.

Acknowledgments: The authors thank David Birdsong and three anonymous reviewers for comments and suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

Notes

[1] $p(H_0|null) = p(null|H_0) * prior(H_0)/p(null) = 0.95 * 0.5/(0.95 * 0.5 + 0.925 * 0.5) = 0.506$.

[2] The estimated effect is 0.20 with 95% confidence interval $[-0.08, 0.48]$ using a random effects model, after excluding Hartshorne et al. (2018). The heterogeneity is larger but still insignificant ($I^2 = 0.44$, $p = 0.08$).

[3] Technically, some of these are simultaneous multilinguals. As noted previously, we use the term 'bilingual' in the more general sense of more than one language.

[4] One might expect, given their more variable circumstances, simultaneous bilinguals might show more variability in performance than do monolinguals. We compared the variability (as measured by coefficient of variation) for monolinguals and simultaneous bilinguals for ages 12 through 67 in year-sized bins (i.e., 7 yo, 8 yo, 9 yo, ... 75 yo), restricted to those who listed English as their primary language. This age range was chosen to ensure at least 20 subjects per bin; less than that, and measurement of coefficient of variation is very unstable. While accuracy was significantly higher for monolinguals relative to simultaneous bilinguals at nearly every age (see HTP), variation was generally similar at each age and did not differ as a group (t(55) = 1.73, p = 0.09). Similar results were obtained when measuring subject performance using expected ability inferred from a 4PL IRT model (cf. Hartshorne and Chen (2021)) instead of using elogit-transformed accuracy (as done here and in HTP). However, we should note that estimating variability is difficult when performance is close to ceiling, which is the case for both monolinguals and simultaneous bilinguals.

[5] These estimates do not necessarily exclude non-significant results. However, since the vast majority of reported results are significant, this is unlikely to bias estimates much (Fanelli 2010; Sterling et al. 1995).

References

Ardila, Alfredo, Mónica Rosselli, Alexandra Ortega, Merike Lang, and Valeria L. Torres. 2019. Oral and written language abilities in young spanish/english bilinguals. *International Journal of Bilingualism* 23: 296–312. [CrossRef]

Bakker, Marjan, Annette Van Dijk, and Jelte M. Wicherts. 2012. The rules of the game called psychological science. *Perspectives on Psychological Science* 7: 543–54. [CrossRef] [PubMed]

Balduzzi, Sara, Gerta Rücker, and Guido Schwarzer. 2019. How to perform a meta-analysis with R: A practical tutorial. *Evidence-Based Mental Health* 22: 153–60. [CrossRef] [PubMed]

Birdsong, David. 2005. Nativelikeness and non-nativelikeness in L2A research1. *IRAL, International Review of Applied Linguistics in Language Teaching* 43: 319. [CrossRef]

Birdsong, David, and Libby M. Gertken. 2013. In faint praise of folly: A critical review of native/non-native speaker comparisons, with examples from native and bilingual processing of french complex syntax. *Language, Interaction and Acquisition* 4: 107–33. [CrossRef]

Bylund, Emanuel, Kenneth Hyltenstam, and Niclas Abrahamsson. 2020. Age of acquisition—not bilingualism—is the primary determinant of less than nativelike L2 ultimate attainment. *Bilingualism: Language and Cognition* 24: 18–30. [CrossRef]

Campbell, Mhairi, Joanne E. McKenzie, Amanda Sowden, Srinivasa Vittal Katikireddi, Sue E. Brennan, Simon Ellis, Jamie Hartmann-Boyce, Rebecca Ryan, Sasha Shepperd, James Thomas, and et al. 2020. Synthesis without meta-analysis (SWiM) in systematic reviews: Reporting guideline. *BMJ* 368: l6890. [CrossRef]

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale: Lawrence Erlbaum Associates, pp. 18–74.

Crain, Stephen. 1991. Language acquisition in the absence of experience. *Behavioral and Brain Sciences* 14: 597–612. [CrossRef]

Fanelli, Daniele. 2010. "Positive" results increase down the hierarchy of the sciences. *PLoS ONE* 5: e10068. [CrossRef]

Fleiss, J. 1993. Review papers: The statistical basis of meta-analysis. *Statistical Methods in Medical Research* 2: 121–45. [CrossRef]

Garraffa, Maria, Mateo Obregon, and Antonella Sorace. 2017. Linguistic and cognitive effects of bilingualism with regional minority languages: A study of sardinian–italian adult speakers. *Frontiers in Psychology* 8: 1907. [CrossRef]

Garraffa, Maria, Mateo Obregon, Bernadette O'Rourke, and Antonella Sorace. 2020. Language and cognition in gaelic-english young adult bilingual speakers: A positive effect of school immersion program on attentional and grammatical skills. *Frontiers in Psychology* 11: 2758. [CrossRef] [PubMed]

Gignac, Gilles E., and Eva T. Szodorai. 2016. Effect size guidelines for individual differences researchers. *Personality and Individual Differences* 102: 74–78. [CrossRef]

Giguere, David, and Erika Hoff. 2020. Home language and societal language skills in second-generation bilingual adults. *International Journal of Bilingualism* 24: 1071–87. [CrossRef]

Grosjean, François. 1989. Neurolinguists, beware! The bilingual is not two monolinguals in one person. *Brain and Language* 36: 3–15. [CrossRef]

Grosjean, François. 2016. The complementarity principle and its impact on processing, acquisition, and dominance. In *Language Dominance in Bilinguals*. Edited by Jeanine Treffers-Daller and Carmen Silva-Corvalán. Cambridge: Cambridge University Press, pp. 66–84. [CrossRef]

Hartshorne, Joshua K., and Adena Schachner. 2012. Tracking replicability as a method of post-publication open evaluation. *Frontiers in Computational Neuroscience* 6: 8. [CrossRef]

Hartshorne, Joshua K., and Laura T. Germine. 2015. When does cognitive functioning peak? The asynchronous rise and fall of different cognitive abilities across the life span. *Psychological Science* 26: 433–43. [CrossRef]

Hartshorne, Joshua K., and Tony Chen. 2021. More evidence from over 1.1 million subjects that the critical period for syntax closes in late adolescence. *Cognition 214*: 104706.

Hartshorne, Joshua K., Joshua B. Tenenbaum, and Steven Pinker. 2018. A critical period for second language acquisition: Evidence from 2/3 million English speakers. *Cognition* 177: 263–77. [CrossRef]

Hemphill, James F. 2003. Interpreting the magnitudes of correlation coefficients. *The American Psychologist* 58: 78–79. [CrossRef]

Hurford, James R. 1991. The evolution of the critical period for language acquisition. *Cognition* 40: 159–201. [CrossRef]

Jaeger, T. Florian. 2008. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language* 59: 434–46. [CrossRef] [PubMed]

Klein, Richard A., Michelangelo Vianello, Fred Hasselman, Byron G. Adams, Reginald B. Adams Jr., Sinan Alper, Mark Aveyard, Jordan R. Axt, Mayowa T. Babalola, Štěpán Bahník, and et al. 2019. Many labs 2: Investigating variation in replicability across sample and setting. *Advances in Methods and Practices in Psychological Science* 1: 443–90. [CrossRef]

Lakshmanan, Usha. 2013. *Bilingual Assessment*. Hoboken: John Wiley & Sons, Inc. [CrossRef]

Langsford, Steven, Amy Perfors, Andrew T. Hendrickson, Lauren A. Kennedy, and Danielle J. Navarro. 2018. Quantifying sentence acceptability measures: Reliability, bias, and variability. *Glossa: A Journal of General Linguistics* 3: 37. [CrossRef]

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349. science.aac4716. [CrossRef]

Page, Matthew J., Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, and et al. 2021. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *International Journal of Surgery* 88: 105906. [CrossRef] [PubMed]

Pinker, Steven. 1994. *The Language Instinct.* New York: William Morrow and Company.

Plonsky, Luke, Ekaterina Sudina, and Yuhang Hu. 2021. Applying meta-analysis to research on bilingualism: An introduction. *Bilingualism: Language and Cognition* 24: 819–24. [CrossRef]

Richard, F. Dan, Charles F. Bond Jr., and Juli J. Stokes-Zoota. 2003. One hundred years of social psychology quantitatively described. *Review of General Psychology* 7: 331–63. [CrossRef]

Schäfer, Thomas, and Marcus A. Schwarz. 2019. The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology* 10: 813. [CrossRef]

Serratrice, Ludovica. 2013. Cross-linguistic influence in bilingual development: Determinants and mechanisms. *Linguistic Approaches to Bilingualism* 3: 3–25. [CrossRef]

Snedeker, Jesse, Joy Geren, and Carissa L. Shafto. 2012. Disentangling the effects of cognitive development and linguistic expertise: A longitudinal study of the acquisition of english in internationally-adopted children. *Cognitive Psychology* 65: 39–76. [CrossRef]

Stanley, Tom D., Evan C. Carter, and Hristos Doucouliagos. 2018. What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin* 144: 1325. [CrossRef] [PubMed]

Sterling, Theodore D., Wilf L. Rosenbaum, and James J. Weinkam. 1995. Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician* 49: 108–12.

Street, James A., and Ewa Dąbrowska. 2010. More individual differences in language attainment: How much do adult native speakers of english know about passives and quantifiers? *Lingua* 120: 2080–94. [CrossRef]

Thordardottir, Elin, Alyssa Rothenberg, Marie-Eve Rivard, and Rebecca Naves. 2006. Bilingual assessment: Can overall proficiency be estimated from separate measurement of two languages? *Journal of Multilingual Communication Disorders* 4: 1–21. [CrossRef]

Wexler, Ken. 1998. Very early parameter setting and the unique checking constraint: A new explanation of the optional infinitive stage. *Lingua 106*: 23–79. [CrossRef]

Zipf, George Kingsley. 1935. *The Psycho-Biology of Language*. Cambridge: The MIT Press.