Bilingualism: Language and Cognition

cambridge.org/bil

Divergence point analyses of visual world data: applications to bilingual research

Kate Stone[1], Sol Lago[2] and Daniel J. Schad[3]

[1]University of Potsdam; [2]Goethe University Frankfurt and [3]University of Tilburg

Review Article

Cite this article: Stone K, Lago S, Schad DJ (2021). Divergence point analyses of visual world data: applications to bilingual research. *Bilingualism: Language and Cognition* **24**, 833–841. https://doi.org/10.1017/S1366728920000607

Received: 7 February 2020
Revised: 18 September 2020
Accepted: 18 September 2020
First published online: 10 December 2020

Keywords:
divergence point analyses; non-parametric approaches; bootstrapping; visual world eye-tracking; bilingualism

Address for correspondence:
Kate Stone, Email: stone@uni-potsdam.de

Abstract

Much work has shown that differences in the timecourse of language processing are central to comparing native (L1) and non-native (L2) speakers. However, estimating the onset of experimental effects in timecourse data presents several statistical problems including multiple comparisons and autocorrelation. We compare several approaches to tackling these problems and illustrate them using an L1-L2 visual world eye-tracking dataset. We then present a bootstrapping procedure that allows not only estimation of an effect onset, but also of a temporal confidence interval around this divergence point. We describe how divergence points can be used to demonstrate timecourse differences between speaker groups or between experimental manipulations, two important issues in evaluating L2 processing accounts. We discuss possible extensions of the bootstrapping procedure, including determining divergence points for individual speakers and correlating them with individual factors like L2 exposure and proficiency. Data and an analysis tutorial are available at https://osf.io/exbmk/.

1. Introduction

Studying the timecourse of comprehension is a central goal in bilingual processing research, which has been significantly fostered by the use of time-sensitive methods such as self-paced reading, eye-tracking, and event-related potentials. The importance of timing is highlighted by findings showing that comprehension is often slower in a non-native than in a native language, in both lexical and sentence domains. For example, compared to monolinguals, even highly proficient bilinguals show slower lexical access (Duyck, Vanderelst, Desmet & Hartsuiker, 2008; Gollan, Slattery, Goldenberg, Van Assche, Duyck & Rayner, 2011; Lehtonen, Hultén, Rodríguez-Fornells, Cunillera, Tuomainen & Laine, 2012; Lemhöfer, Spalek & Schriefers, 2008; Ransdell & Fischler, 1987). Similarly, sentence processing studies often find that when a word violates a grammatical constraint or a previously established parse, monolinguals display processing disruptions soon after the violation, while disruptions in bilinguals are often delayed (Boxell & Felser, 2017; Felser & Cunnings, 2012; Grüter, Lew-Williams & Fernald, 2012; Hopp, 2017; Steinhauer, White & Drury, 2009; White, Genesee & Steinhauer, 2012).

Despite the rich data generated by current methods, our inferences about L1-L2 temporal asymmetries are often limited by using methods demonstrating that differences in native vs. non-native processing affect different sentence regions (in self-paced reading), different temporal windows (in event-related potentials), or different reading measures (in eye-tracking). Instead, it would be preferable to establish the precise timepoint at which an effect onsets in order to directly compare timing differences between speaker groups or between experimental manipulations. This article summarizes several techniques for achieving this goal. Such information is relevant to testing a variety of L2 accounts. For example, some accounts propose that L1-L2 processing differences concern the relative timing of grammatical versus non-grammatical information (Clahsen & Felser, 2018). Meanwhile, capacity-based accounts link timing delays to differential proficiency, lexical access speed, and working memory (Dekydtspotter & Renaud, 2014; Hopp, 2013; McDonald, 2006). Establishing numeric divergence points rather than dichotomous contrasts (effect present/absent) would allow testing of whether timing delays are predicted by such variables.

To encourage the use of divergence point analyses, we provide a practical introduction using a L1-L2 visual world eye-tracking dataset. The data and a step-by-step R analysis tutorial are available at https://osf.io/exbmk/. Finally, we note that divergence point analyses differ from another set of techniques which examine timeseries data by modeling the shape (i.e., functional form) of change across time (e.g., Mirman, 2017; Porretta, Kyröläinen, van Rij & Järvikivi, 2018). When characterizing timeseries data, both types of techniques are useful and provide complementary information.

CAMBRIDGE UNIVERSITY PRESS

2. A practical example

Our L1-L2 dataset belongs to a visual world experiment examining the use of syntactic gender information to make noun predictions. The visual world paradigm involves tracking eye movements to objects on a computer screen while participants hear a sentence, with the assumption that there is a close link between eye movements and language processes (Huettig, Rommers & Meyer, 2011). The visual world paradigm is thus particularly useful in L1-L2 timecourse research because it measures how language processing unfolds over time.

We tested a group of L1 German speakers and two groups of intermediate-to-advanced L2 German speakers, whose L1 was either Spanish or English (for demographic details see Appendix S1, Supplementary Materials). Participants saw four objects on a computer display and heard a German instruction to click on one of the objects as quickly as possible, e.g., *Click on the blue button* (Figure 1A). The determiner and adjective in the instruction agreed in gender and color with only one of the objects (henceforth the "target"), allowing participants to identify it prior to its pronunciation; namely, at the adjective (Hopp & Lemmerth, 2018; Lemmerth & Hopp, 2019). The properties of the other objects were manipulated such that they matched the target only in color ("color competitor"), only in gender ("gender competitor"), or neither ("distractor").

The critical time window for assessing gender predictions was from the onset of the adjective to 200 ms after the onset of the noun, to account for the time taken to program and launch an eye movement (Hallett, 1986; Salverda, Kleinschmidt & Tanenhaus, 2014). As Figure 1B shows, fixations before the adjective were distributed similarly between the four objects. At the adjective, fixations to the target and the color competitor increased, while looks to the gender competitor and distractor abruptly decayed. Given this pattern, we focused on the divergence between the target and color competitor (henceforth, "competitor"). As both objects match the color of the adjective, but only the target has the appropriate gender, any target-over-competitor advantage should reflect the predictive use of gender.

We used a divergence point analysis to establish how soon after the adjective a target-over-competitor advantage appeared (i.e., a predictive effect), and whether this divergence occurred later in L2 than L1 speakers, consistent with previous findings (Dussias, Valdés Kroff, Guzzardo Tamargo & Gerfen, 2013; Grüter et al., 2012; Hopp, 2013; Lew-Williams & Fernald, 2010). We also wanted to determine whether gender predictions were modulated by participants' native language. If so, Spanish speakers may benefit from the rich morphosyntactic gender agreement of their L1 and show faster predictions than English speakers, whose L1 lacks syntactic gender agreement.

3. Divergence point analyses: an intuitive approach

One possible approach to determine the divergence point between looks to the target vs. competitor is to statistically compare the difference in fixation proportions at each timepoint and find the earliest significant test statistic. To illustrate this, we use our data where eye positions were sampled at 50 Hz, i.e., every 20 ms. At each sampled timepoint, we fit a generalized logistic mixed-effects model with a binomial distribution to compare the proportion of fixations to the target vs. competitor (GLMM; Barr, 2008). The divergence point was defined as the earliest point with a significant positive estimate (Figure 2).

Although this approach is intuitive, it involves as many statistical comparisons as there are timepoints and thus runs a risk of false positives (Type 1 error). For example, at an alpha of 0.05, the probability of a single test delivering a false positive is 5%. But with 45 timepoints in our window of interest, this probability rises to 90% ($1 - 0.95^{45}$). The combined probability of a false positive over an entire set of tests is known as the family-wise error rate (FWER; Hochberg & Tamhane, 1987).

A common way to control for multiple comparisons is the Bonferroni correction, which lowers the alpha-level by dividing the desired alpha by the number of tests (Bonferroni, 1936). Thus, the alpha-level for 45 tests becomes 0.001 and the FWER at this adjusted alpha is around 5%. The downside of the Bonferroni correction is that lowering alpha necessarily decreases statistical power, because it becomes more difficult for an effect (true or otherwise) to reach the significance threshold. Thus, the larger the number of tests, the lower the power to detect a true effect.

A second type of correction that preserves power is false discovery rate (FDR) control (Benjamini & Hochberg, 1995). Instead of correcting the alpha level, FDR control restricts the proportion of false discoveries among the significant results. To apply FDR control, we take the p-values from the 45 tests and sort them from smallest to largest. A critical value for each p-value is then calculated via a suitable method (e.g., some methods account for autocorrelated data, others for data with many significant results; Benjamini, Krieger & Yekutieli, 2006; Benjamini & Yekutieli, 2001). The largest p-value below this critical value is then chosen as the new significance cut-off for the original p-values.

Both the Bonferroni correction and FDR control can be easily implemented (**code §2**). Figure 2 shows the corrected and uncorrected divergence point estimates for our data. As expected, corrected estimates are always later than uncorrected estimates, suggesting that the latter are false positives. The higher power of FDR control over Bonferroni is visible in the German and English groups; whereas in the Spanish group, where the difference in fixation proportions arises more abruptly, both corrections yield similar results.

While the corrections account for FWER, an additional issue in visual world data is autocorrelation. Autocorrelation occurs because modern eye-trackers can record eye fixations at high frequencies (e.g., once per millisecond), but planning and executing an eye movement takes around 200 milliseconds. Thus, neighboring datapoints often reflect the same stage of cognitive processing and so are strongly correlated. Applying parametric tests at multiple timepoints will overestimate variance in the data and can influence the Type 1 error rate, because parametric tests assume independent observations. Importantly, grouping observations into larger bins can reduce (Mirman, 2014:18), but not eliminate autocorrelation; Figure 3, **code §3**).

A second dependency issue in our data is the contingency of fixations to the target and competitor, since a participant cannot simultaneously look at both objects. For the same reasons as autocorrelation, this can inflate the Type 1 error rate. Finally, the approach above does not estimate the temporal uncertainty around a divergence point because the latter is based on a single statistical test. While we can estimate the 95% confidence interval of a test coefficient, this reflects uncertainty about the MAGNITUDE of the target vs. competitor difference, rather than the temporal location of the divergence point. In order to statistically compare the onset of predictions between groups, a measure of temporal variability is necessary. With this goal, we turn to non-parametric resampling approaches.

https://doi.org/10.1017/S1366728920000607 Published online by Cambridge University Press

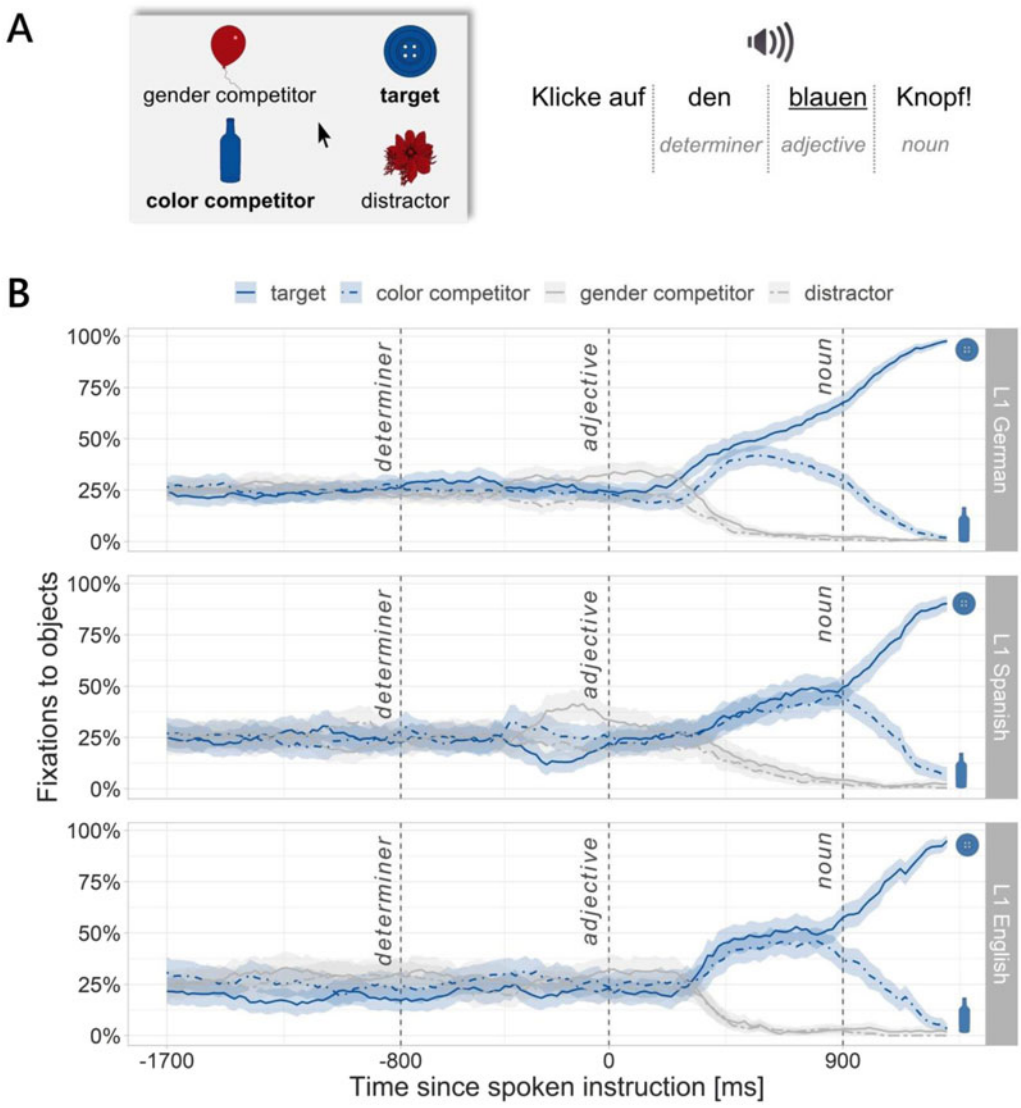

Fig. 1. (A) Sample visual display and auditory instruction (translation: '*Click on the.*$_{MASC}$ *blue.*$_{MASC}$ *button.*$_{MASC}$'). Only the target object matched the gender and color cues of the determiner and adjective. The other three objects matched the target only in color (bottle$_{.FEM}$: *color competitor*), only in gender (balloon$_{.MASC}$: *gender competitor*), or neither (flower$_{.FEM}$: *distractor*). **(B)** Percentage of fixations to the four objects in each speaker group. Lines show mean fixation percentages and shading shows 95% bootstrapped confidence intervals. The onset of the target noun is displayed 200 ms shifted to the right.

4. Non-parametric approaches

The corrected comparisons above allow us to estimate a divergence point indexing the onset of predictive looks. But how certain are we about this estimate? Because we only conducted our procedure once, we cannot be sure that a similar divergence point would be found in a different sample. Non-parametric approaches such as bootstrapping and cluster permutation can answer this question by resampling or permuting existing data to generate "new" datasets and sampling the distributions of their test statistics. Conveniently, they also control for FWER and autocorrelation (Groppe, Urbach & Kutas, 2011; Maris & Oostenveld, 2007; Reingold & Sheridan, 2014).

4.1 Cluster permutation tests

Cluster permutation identifies temporal "clusters" in which two experimental conditions differ (Barr, Jackson & Phillips, 2014). In our dataset, these clusters would represent time windows in which looks to the target and competitor differed significantly. In a permutation test, condition labels (e.g., target/competitor) are randomly reassigned multiple times in order to destructure the experimental manipulation and generate a distribution of test results consistent with the null hypothesis. The significance of the test statistic from the original dataset is then based on its relative position in the permutation-derived null distribution. FWER is controlled by reducing the number of statistical comparisons to one. Autocorrelation is also controlled, because the temporal structure of the data is preserved during permutation. Thus, the effect of autocorrelation is constant across permutations and the only factor affecting the variance of the permutation distribution is the reassignment of condition labels.

However, one disadvantage of cluster-based permutation is that significant clusters do not indicate when an effect arose or its temporal variability, but rather only that there was a window in which an effect was significant (Maris & Oostenveld, 2007; Sassenhagen & Draschkow, 2019). Since our research question

https://doi.org/10.1017/S1366728920000607 Published online by Cambridge University Press

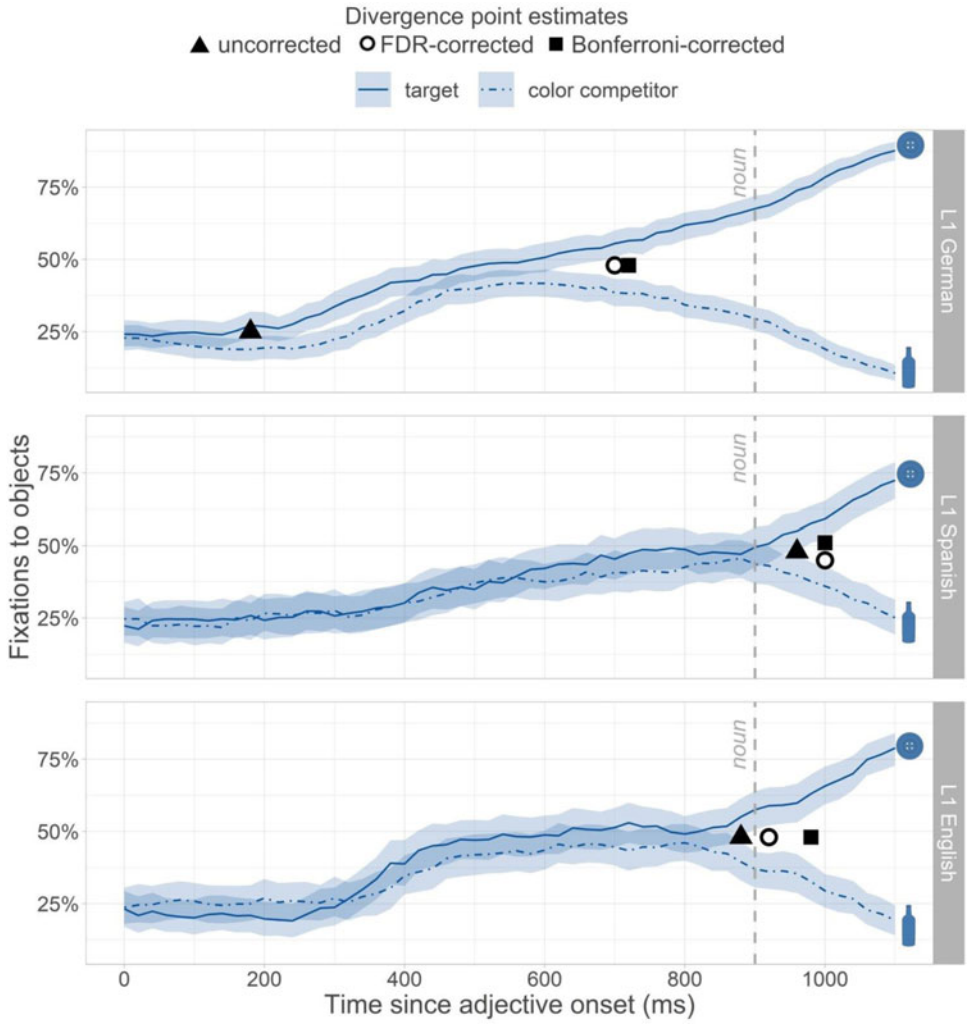

Fig. 2. Estimated onset of predictive looks to the target vs. competitor using GLMM tests at each timepoint with either no correction for multiple comparisons, a Bonferroni correction, or false discovery rate (FDR) control. Both corrections result in later, more conservative divergence point estimates relative to uncorrected estimates.

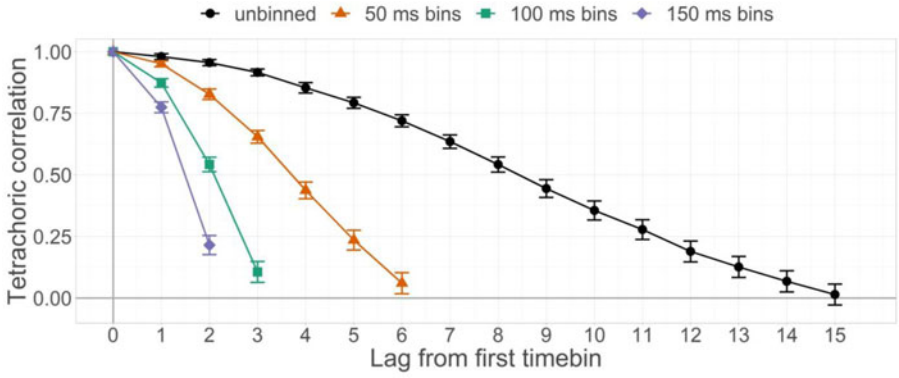

Fig. 3. Tetrachoric correlations of target fixation probabilities between each timebin and the first bin of the series, plotted as a function of bin size. The "unbinned" black line reflects the correlation between fixations sampled every 20 ms. Error bars indicate standard errors. A correlation of 1 at a 0-lag indicates the correlation of a bin with itself. As the lag increases, autocorrelation decreases. The plot demonstrates that most correlations are not consistent with zero. Even large bins do not completely eliminate autocorrelation between bins and come at the expense of temporal precision.

concerns the onset of predictive looks, below we demonstrate an approach that can address this question while preserving the advantages of non-parametric approaches.

4.2 Bootstrapping

The goal of bootstrapping is to estimate what the distribution of statistical test results would be if we repeated our experiment many times. For this, an existing dataset is resampled multiple times to generate "new" datasets and a statistical test is applied after each resample. Below we use a NON-PARAMETRIC bootstrap, which does not make assumptions about the population distribution underlying the data, meaning that it can be used for non-normally distributed data (Maris & Oostenveld, 2007; Hesterberg, 2002). The bootstrapping technique has previously been applied to reading eye-tracking and event-related potentials

https://doi.org/10.1017/S1366728920000607 Published online by Cambridge University Press

(Schad, Risse, Slattery & Rayner, 2014; Wasserman & Bockenholt, 1989; Sheridan & Reingold, 2012; Reingold & Sheridan, 2014), and its results have been shown to be comparable to those of permutation tests (Rosenfeld & Donchin, 2015). A bootstrapping approach for visual world data is presented in Seedorff, Oleson and McMurray (2018), although it answers a different research question from the one of interest here.

The steps in our approach are as follows. First, for each speaker group, we extract data where either the target or competitor was fixated. To identify a divergence point between fixations, we apply an uncorrected statistical test at each timepoint aggregating over items (**code §4**). Here we use a one-sample t-test on fixation proportions because it is conceptually straightforward and convenient in terms of convergence and computational time. T-tests are often used in non-parametric methods (e.g., Groppe et al., 2011; Maris & Oostenveld, 2007; Efron & Tibshirani, 1986; Hesterberg, 2015; Reingold & Sheridan, 2014). For data that do not include a large number of extreme values (e.g., clustered close to 0% or 100%), a t-test reasonably approximates the results of a logistic model, which would be a more appropriate choice given the binary nature of our data. However, fitting multiple logistic models with the appropriate random effects structure

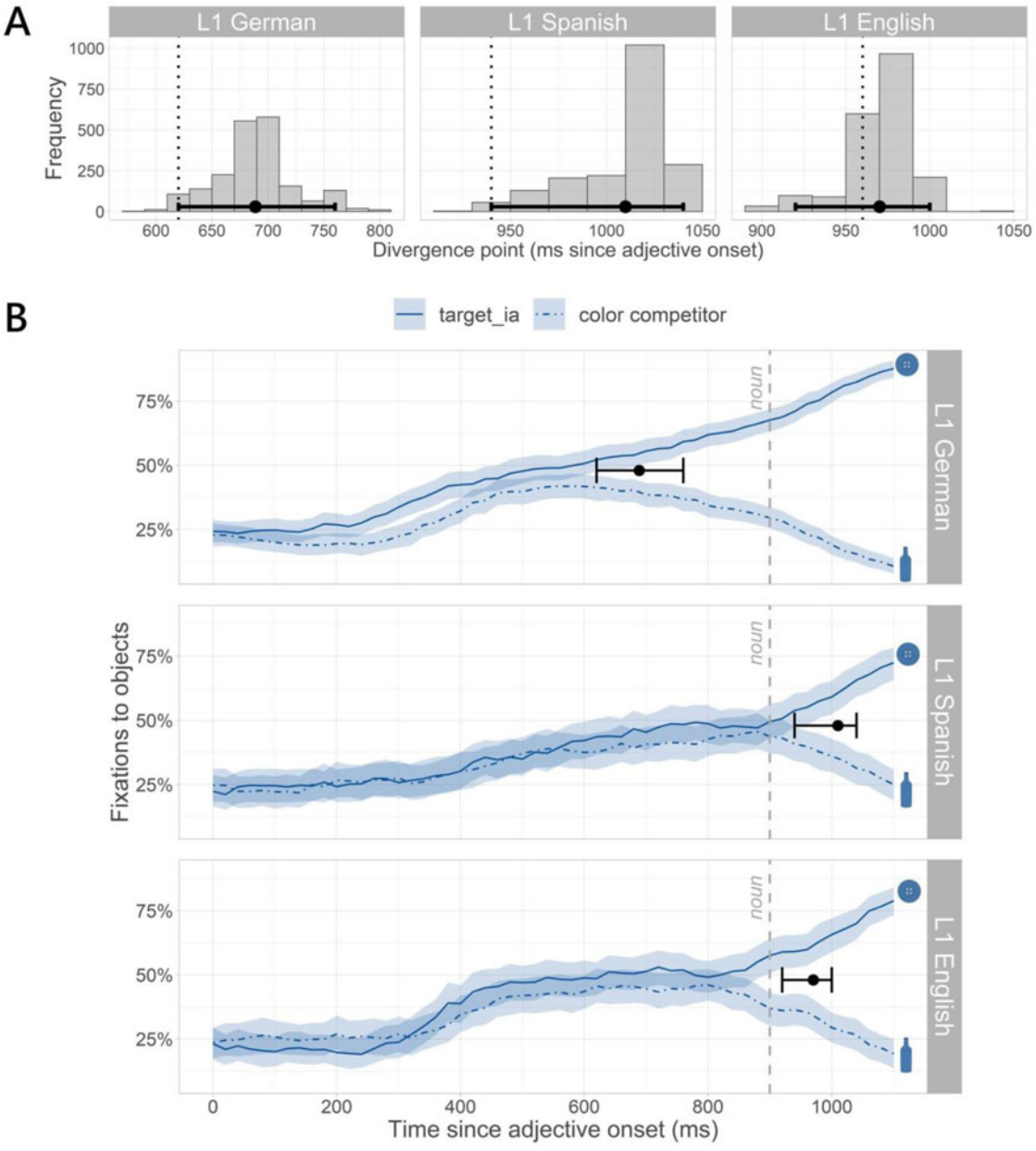

Fig. 4. (A) Bootstrap distributions of divergence points for each language group. The x-axis shows the distribution of divergence points based on 2000 bootstraps. The y-axis shows the number of resamples where a given divergence point was observed. Points with error bars indicate the bootstrap mean and its 95% percentile confidence interval, which reflect divergence points and their temporal uncertainty. Dotted vertical lines represent the divergence points in the original data. The difference between the empirical and bootstrap means, or bias, is used as a diagnostic of the bootstrap's ability to recover the mean of the population—which is assumed to be represented by the mean of the original sample. **(B)** Divergence points and 95% confidence intervals superimposed on the fixation curves. German L1 speakers show the earliest predictive onsets at 689 [620, 760] ms post-adjective. The L2 groups do not appear to predict the target object, as their mean divergence point estimates are after the noun: L1 Spanish speakers 1010 [940, 1040] ms and L1 English speakers 970 [920, 1000] ms.

https://doi.org/10.1017/S1366728920000607 Published online by Cambridge University Press

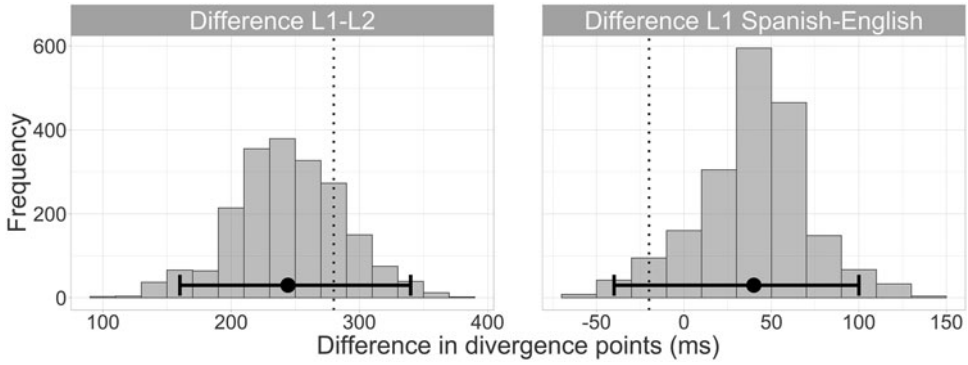

Fig. 5. Bootstrap distributions of the difference in divergence points between L1-L2 speakers (left) and L1 Spanish-English speakers (right). Points and error bars indicate bootstrap means and 95% percentile confidence intervals. Dotted vertical lines indicate mean divergence point differences in the original data. *L1-L2 comparison*: divergence point difference = 244 ms, 95% CI = [160, 340] ms. *L1 Spanish-English comparison*: divergence point difference = 40 ms, 95% CI = [−40, 100] ms.

comes at the expense of increased complexity and computation time. For a comparison between different tests see Appendix S2 (Supplementary Materials).

To establish a divergence point, we take the first timepoint in a run of at least 10 consecutive timepoints with significant t-values. A run of 10 is used because we are interested in the beginning of SUSTAINED looks to the target (in our case, at least 200 ms given the 50 Hz sampling rate). Researchers should choose their own threshold depending on their research question and experimental design.

Next, we use a non-parametric bootstrap to generate "new" datasets by resampling the original dataset with replacement. The resampling is stratified by participant, timepoint, and object type (target/competitor), meaning that data are resampled within these categories. A new divergence point is estimated after each resample. With sufficient resampling (1000–2000 times; Efron & Tibshirani, 1993) a distribution of divergence points is generated whose mean is taken as the overall divergence point (Figure 4A). Variability around the mean of the bootstrap distribution can be quantified with a confidence interval (CI), calculated via a method suited to the properties of the bootstrap distribution and computation time (Carpenter & Bithell, 2000; DiCiccio & Efron, 1996).

Bootstrapped means and CIs for each group are plotted in Figure 4. To compare between groups, we can bootstrap the difference between their divergence points. The result is a distribution of differences (Figure 5; **code §4.4**). The mean difference in divergence points between the L1 and L2 groups is 244 ms, 95% CI = [160, 340] ms. The CI does not contain zero and thus supports a reliable difference. Between the Spanish and English groups, the mean difference is 40 ms, 95% CI = [−40, 100] ms, consistent with a slightly earlier divergence point in the English group. However, the CI of the between-group difference contains zero and thus fails to support a difference. If desired, p-values can also be computed (see **Appendix S4,** Supplementary Materials).

The results show that L2 speakers are slower than native speakers to start looking preferentially at the target, consistent with a predictive advantage in the native speaker group. Further, both L2 groups show mean divergence points after the appearance of the noun, which is not consistent with a predictive use of gender. The lack of evidence for an earlier onset in Spanish vs. English speakers does not support the claim that having a gendered native language enhances its predictive use in a foreign language. Instead, it is consistent with a general delay due to L2 status. Note that studies relying on time-window analyses would have reached a similar conclusion by showing significant effects in earlier time windows for L1 than L2 speakers (e.g., by stating that an effect is significant in one group but absent in another). The critical contribution of the bootstrapping method is that it precisely quantifies the delay in the L2 speakers, while allowing a direct between-group comparison of divergence points and estimating their uncertainty.

4.3 Advantages and disadvantages of the bootstrapping approach

Above we demonstrate that resampling approaches can control FWER and autocorrelation in time series analyses. The main advantage of our bootstrapping approach is that it quantifies divergence points and their temporal uncertainty, enabling statistical comparisons between participant groups and/or experimental conditions. However, one disadvantage of the approach is that it does not estimate the duration of an effect or the presence of multiple divergences, although it could be extended to do so. Second, our approach – and onset detection approaches in general – may not be appropriate for analyses where the research question concerns WHETHER an effect is present (Seedorff et al., 2018). Our approach assumes that an effect is present and that the task is simply to detect its onset.

Furthermore, resampling approaches like bootstrapping can describe a dataset but are not generative models. Generative models provide explicit assumptions to connect data with cognitive processes of interest, allowing researchers to examine the parameters that best explain the data and to compare the goodness of fit of different models (Vandekerckhove, Matzke & Wagenmakers, 2015). Two generative approaches that allow divergence point estimation include generalized additive mixed-effects models (GAMMs; van Rij, 2015; van Rij, Vaci, Wurm & Feldman, 2020; Miwa & Baayen, 2020) and Bootstrapped Differences of Timeseries (BDOTS; Seedorff et al., 2018). GAMMs are regression models that estimate non-linear patterns from timecourse data (**Appendix S3,** Supplementary Materials). BDOTS fits 4-parameter logistic and double Gaussian functions to individual fixation curves, which are then bootstrapped to estimate the standard error of mean fixations at each time point in the series. The onset of a divergence in fixations between conditions is then established via t-tests and a Bonferroni correction modified to account for autocorrelation.

https://doi.org/10.1017/S1366728920000607 Published online by Cambridge University Press

Table 1. Comparison of methods suitable for timecourse analysis

Method	Provides divergence point estimates?	Generative model?	Detects (✓) vs. assumes (×) effect?	Estimates uncertainty around a divergence point?	Can divergence points be statistically compared?
Bootstrapping	✓	×	×	✓	✓
Cluster permutation	×	×	✓	×	×
BDOTS	✓	✓	✓	×	×
GAMMs	✓	✓	✓	×	×

Note. Bootstrapping refers to our proposed approach; BDOTS stands for Bootstrapped Differences of Timeseries (Seedorff et al., 2018) and GAMMs for Generalized Additive Mixed-Effect Models (van Rij, 2015; van Rij et al., 2020; Miwa & Baayen, 2020).

The downside of models such as GAMMs and BDOTS is that they do not provide a measure of variability around their divergence point estimates, which is needed for statistical comparison (Table 1). Furthermore, while GAMMs can estimate a within-condition divergence point, BDOTS can only estimate a divergence point between conditions.

5. Further applications to bilingualism

Onset estimates can enrich L2 processing theories in several ways. Consider, for example, the claim made by capacity-based accounts that processing is slower in L2 than L1 due to limits on lexical access speed and working-memory capacity (Dekydtspotter & Renaud, 2014; Hopp, 2013; McDonald, 2006). These two constructs can already be measured quantitatively using word recognition and working-memory span tasks, but it is unclear how well they predict processing speed during sentence comprehension. Having precise estimates of prediction speed would allow us to answer this question and provide a more precise evaluation of capacity-based accounts.

Another useful application concerns L2 accounts that posit that non-native and native speakers weigh different kinds of information differently in processing (Clahsen & Felser, 2018; Cunnings, 2017). Some of this research has found that L2 speakers are often slower (or less sensitive) than native speakers to syntactic information, but more sensitive to discourse-level information like extra-sentential context and semantic plausibility (Felser & Cunnings, 2012; Pan, Schimke & Felser, 2015; Roberts & Felser, 2011). Having a method to formally establish when different information sources affect L2 processing would provide key data to test these claims.

Finally, our bootstrapping method can be adapted to quantify variability between speakers. For example, our failure to find Spanish vs. English group differences may have resulted from our sample demographic properties (e.g., potential differences in L2 age of acquisition or proficiency). While data from individual participants is noisier than averaged data, the analysis presented here could be performed on a by-participant basis given a sufficient number of trials (Reingold & Sheridan, 2014), allowing us to examine correlation between individual divergence points and factors like proficiency and L2 exposure. Together with the estimation of by-group timing effects, we believe that quantifying individual variability will prove crucial to improve models of bilingual processing.

Supplementary Material. For supplementary material accompanying this paper, visit https://doi.org/10.1017/S1366728920000607

List of supplementary materials:

S1. Demographic profiles of the language groups.

S2. Comparison of different statistical tests in the bootstrap approach.

S3. Comparison of bootstrap- and GAMM-derived divergence points.

S4. Null hypothesis tests of the bootstrapped estimates

Acknowledgements. We thank Harald Baayen, Pantelis Bagos, Dale Barr, Jason Geller, Ben Goodrich, Carrie Neal Jackson, Lauren Kennedy, Dorothea Pregla, João Veríssimo, and Titus von der Malsburg for valuable comments and feedback. Kate Stone and Sol Lago were supported by a German Research Council project awarded to Sol Lago (grant number LA 3774/1-1).We also thank Lisa Becker for her assistance in building the experiment and collecting data.

References

Agresti A (2003) *Categorical Data Analysis*. John Wiley & Sons.

Barr DJ (2008) Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language* **59**, 457–474. https://doi.org/10.1016/j.jml.2007.09.002

Barr DJ, Jackson L and Phillips I (2014) Using a voice to put a name to a face: The psycholinguistics of proper name comprehension. *Journal of Experimental Psychology: General* **143**, 404–413. https://doi.org/10.1037/a0031813

Barr DJ, Levy R, Scheepers C and Tily HJ (2013) Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* **68**, 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates D, Kliegl R, Vasishth S and Baayen H (2018) Parsimonious Mixed Models. *ArXiv:*1506.04967 *[Stat]*. http://arxiv.org/abs/1506.04967

Benjamini Y and Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Benjamini Y, Krieger AM and Yekutieli D (2006) Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* **93**, 491–507. https://doi.org/10.1093/biomet/93.3.491

Benjamini Y and Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* **29**, 1165–1188. https://doi.org/10.1214/aos/1013699998

Bonferroni CE (1936) Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni Del R Istituto Superiore Di Scienze Economiche e Commerciali Di Firenze.*, **8**, 3–62.

Boxell O and Felser C (2017) Sensitivity to parasitic gaps inside subject islands in native and non-native sentence processing*. *Bilingualism: Language and Cognition* **20**, 494–511. https://doi.org/10.1017/S1366728915000942

Carpenter J and Bithell J (2000) Bootstrap confidence intervals: When, which, what? A practical guide for medical statisticians. *Statistics in Medicine* **19**, 1141–1164.

Clahsen H and Felser C (2018) Some notes on the Shallow Structure Hypothesis. *Studies in Second Language Acquisition* **40**, 693–706. https://doi.org/10.1017/S0272263117000250

Cunnings I (2017) Parsing and Working Memory in Bilingual Sentence Processing. *Bilingualism: Language and Cognition* **20**, 659–678. https://doi.org/10.1017/S1366728916000675

Dekydtspotter L and Renaud C (2014) On second language processing and grammatical development: The parser in second language acquisition.

Linguistic Approaches to Bilingualism **4**, 131–165. https://doi.org/10.1075/lab.4.2.01dek

DiCiccio TJ and Efron B (1996) Bootstrap Confidence Intervals. *Statistical Science* **11**, 189–212. JSTOR.

Donnelly S and Verkuilen J (2017) Empirical logit analysis is not logistic regression. *Journal of Memory and Language* **94**, 28–42. https://doi.org/10.1016/j.jml.2016.10.005

Dussias PE, Valdés Kroff JR, Guzzardo Tamargo RE and Gerfen C (2013) When gender and looking go hand in hand. *Studies in Second Language Acquisition* **35**, 353–387. https://doi.org/10.1017/S0272263112000915

Duyck W, Vanderelst D, Desmet T and Hartsuiker RJ (2008) The frequency effect in second-language visual word recognition. *Psychonomic Bulletin & Review* **15**, 850–855. https://doi.org/10.3758/PBR.15.4.850

Efron B and Tibshirani R (1986) Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science* **1**, 54–75. JSTOR.

Efron B and Tibshirani R (1993) *An introduction to the bootstrap*. CRC Press.

Felser C and Cunnings I (2012) Processing reflexives in a second language: The timing of structural and discourse-level constraints. *Applied Psycholinguistics* **33**, 571–603. https://doi.org/10.1017/S0142716411000488

Gart JJ and Zweifel JR (1967) On the Bias of Various Estimators of the Logit and Its Variance with Application to Quantal Bioassay. *Biometrika* **54**, 181–187. https://doi.org/10.2307/2333861

Gelman A and Hill J (2006) *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gollan TH, Slattery TJ, Goldenberg D, Van Assche E, Duyck W and Rayner K (2011) Frequency drives lexical access in reading but not in speaking: The frequency-lag hypothesis. *Journal of Experimental Psychology: General* **140**, 186–209. https://doi.org/10.1037/a0022256

Groppe DM, Urbach TP and Kutas M (2011) Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review. *Psychophysiology* **48**, 1711–1725. https://doi.org/10.1111/j.1469-8986.2011.01273.x

Grüter T, Lew-Williams C and Fernald A (2012) Grammatical gender in L2: A production or a real-time processing problem? *Second Language Research* **28**, 191–215. https://doi.org/10.1177/0267658312437990

Hallett P (1986) Eye movements and human visual perception. *Handbook of Perception and Human Performance*., **1**, 10–11.

Hesterberg TC (2002) Bootstrap Methods and Permutation Tests. In Moore D, McCabe G, Duckworth W and Sclove S (eds), *The Practice of Business Statistics*. W. H. Freeman.

Hesterberg TC (2015) What Teachers Should Know About the Bootstrap: Resampling in the Undergraduate Statistics Curriculum. *The American Statistician* **69**, 371–386. https://doi.org/10.1080/00031305.2015.1089789

Hochberg Y and Tamhane AC (1987) *Multiple comparison procedures*. John Wiley & Sons.

Hopp H (2013) Grammatical gender in adult L2 acquisition: Relations between lexical and syntactic variability. *Second Language Research* **29**, 33–56. https://doi.org/10.1177/0267658312461803

Hopp H (2017) Individual differences in L2 parsing and lexical representations. *Bilingualism: Language and Cognition* **20**, 689–690. https://doi.org/10.1017/S1366728916000821

Hopp H and Lemmerth N (2018) Lexical and syntactic congruency in L2 predictive gender processing. *Studies in Second Language Acquisition* **40**, 171–199. https://doi.org/10.1017/S0272263116000437

Huettig F, Rommers J and Meyer AS (2011) Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica* **137**, 151–171. https://doi.org/10.1016/j.actpsy.2010.11.003

Lehtonen M, Hultén A, Rodríguez-Fornells A, Cunillera T, Tuomainen J and Laine M (2012) Differences in word recognition between early bilinguals and monolinguals: Behavioral and ERP evidence. *Neuropsychologia* **50**, 1362–1371. https://doi.org/10.1016/j.neuropsychologia.2012.02.021

Lemhöfer K, Spalek K and Schriefers H (2008) Cross-language effects of grammatical gender in bilingual word recognition and production. *Journal of Memory and Language* **59**, 312–330. https://doi.org/10.1016/j.jml.2008.06.005

Lemmerth N and Hopp H (2019) Gender processing in simultaneous and successive bilingual children: Cross-linguistic lexical and syntactic influences. *Language Acquisition* **26**, 21–45. https://doi.org/10.1080/10489223.2017.1391815

Lew-Williams C and Fernald A (2010) Real-time processing of gender-marked articles by native and non-native Spanish speakers. *Journal of Memory and Language* **63**, 447–464. https://doi.org/10.1016/j.jml.2010.07.003

Maris E and Oostenveld R (2007) Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods* **164**, 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Matuschek H, Kliegl R, Vasishth S, Baayen H and Bates D (2017) Balancing Type I Error and Power in Linear Mixed Models. *Journal of Memory and Language* **94**, 305–315.

McCullagh P and Nelder JA (1989) Monographs on statistics and applied probability. *Generalized Linear Models*, 37.

McDonald JL (2006) Beyond the critical period: Processing-based explanations for poor grammaticality judgment performance by late second language learners. *Journal of Memory and Language* **55**, 381–401. https://doi.org/10.1016/j.jml.2006.06.006

Mirman D (2014) *Growth curve analysis and visualization using R*. Chapman and Hall/CRC.

Miwa K and Baayen H (2020) Nonlinearities in bilingual visual word recognition: An introduction to generalized additive modeling. *Bilingualism: Language and Cognition*. Submitted for publication.

Pan H.-Y., Schimke S and Felser C (2015) Referential context effects in non-native relative clause ambiguity resolution. *International Journal of Bilingualism* **19**, 298–313. https://doi.org/10.1177/1367006913515769

Porretta V, Kyröläinen A-J, van Rij J and Järvikivi J (2018) Visual World Paradigm Data: From Preprocessing to Nonlinear Time-Course Analysis. In Czarnowski I, Howlett RJ and Jain LC (eds), *Intelligent Decision Technologies 2017*. Springer International Publishing, pp. 268–277. https://doi.org/10.1007/978-3-319-59424-8_25

Ransdell SE and Fischler I (1987) Memory in a monolingual mode: When are bilinguals at a disadvantage? *Journal of Memory and Language* **26**, 392–405. https://doi.org/10.1016/0749-596X(87)90098-2

Reingold EM and Sheridan H (2014) Estimating the divergence point: A novel distributional analysis procedure for determining the onset of the influence of experimental variables. *Frontiers in Psychology* **5**. https://doi.org/10.3389/fpsyg.2014.01432

Roberts L and Felser C (2011) Plausibility and recovery from garden paths in second language sentence processing. *Applied Psycholinguistics* **32**, 299–331. https://doi.org/10.1017/S0142716410000421

Rosenfeld JP and Donchin E (2015) Resampling (bootstrapping) the mean: A definite do. *Psychophysiology* **52**, 969–972. https://doi.org/10.1111/psyp.12421

Salverda AP, Kleinschmidt D and Tanenhaus MK (2014) Immediate effects of anticipatory coarticulation in spoken-word recognition. *Journal of Memory and Language* **71**, 145–163. https://doi.org/10.1016/j.jml.2013.11.002

Sassenhagen J and Draschkow D (2019) Cluster-based permutation tests of MEG/EEG data do not establish significance of effect latency or location. *Psychophysiology* **56**, e13335. https://doi.org/10.1111/psyp.13335

Schad DJ, Risse S, Slattery T and Rayner K (2014) Word frequency in fast priming: Evidence for immediate cognitive control of eye movements during reading. *Visual Cognition* **22**, 390–414. https://doi.org/10.1080/13506285.2014.892041

Seedorff M, Oleson J and McMurray B (2018) Detecting when timeseries differ: Using the Bootstrapped Differences of Timeseries (BDOTS) to analyze Visual World Paradigm data (and more). *Journal of Memory and Language* **102**, 55–67. https://doi.org/10.1016/j.jml.2018.05.004

Sheridan H and Reingold EM (2012) The time course of predictability effects in reading: Evidence from a survival analysis of fixation durations. *Visual Cognition* **20**, 733–745. https://doi.org/10.1080/13506285.2012.693548

Steinhauer K, White EJ and Drury JE (2009) Temporal dynamics of late second language acquisition: Evidence from event-related brain potentials. *Second Language Research* **25**, 13–41. https://doi.org/10.1177/0267658308098995

van Rij J (2015) *Overview GAMM analysis of time series data*. https://jacolienvanrij.com/Tutorials/GAMM.html#plots

van Rij J, Vaci N, Wurm LH and Feldman LB (2020) Alternative quantitative methods in psycholinguistics: Implications for theory and design. In *Word Knowledge and Word Usage*. De Gruyter Mouton, pp. 83–126. https://doi.org/10.1515/9783110440577-003

Vandekerckhove J, Matzke D and Wagenmakers E.-J. (2015) In Busemeyer JR, Wang Z, Townsend JT and Eidels A (eds), Vol. 1. *Model Comparison and the Principle of Parsimony*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199957996.013.14

Veríssimo J and Clahsen H (2014) Variables and similarity in linguistic generalization: Evidence from inflectional classes in Portuguese. *Journal of Memory and Language* **76**, 61–79. https://doi.org/10.1016/j.jml.2014.06.001

Wasserman S and Bockenholt U (1989) Bootstrapping: Applications to psychophysiology. *Psychophysiology* **26**, 208–221. https://doi.org/10.1111/j.1469-8986.1989.tb03159.x

White EJ, Genesee F and Steinhauer K (2012) Brain responses before and after intensive second language learning: Proficiency based changes and first language background effects in adult learners. *PLOS ONE* **7**, e52318. https://doi.org/10.1371/journal.pone.0052318