Journal of Experimental Psychology:
Learning, Memory, and Cognition
2016, Vol. 42, No. 12, 1994–2001

0278-7393/16/$12.00 http://dx.doi.org/10.1037/xlm0000261

COMMENTARY

Absence of Sublexical Representations in Late-Learning Signers? A Statistical Critique of Lieberman et al. (2015)

Anne Pier Salverda
University of Rochester

Lieberman, Borovsky, Hatrak, and Mayberry (2015) used a modified version of the visual-world paradigm to examine the real-time processing of signs in American Sign Language. They examined the activation of phonological and semantic competitors in native signers and late-learning signers and concluded that their results provide evidence that the mental lexicon of late learners is organized differently from that of native signers. In particular, they claimed that late-learning signers, in contrast to native signers, do not activate phonological competitors during the real-time recognition of spoken words. I argue that this claim receives no substantive support from the data and the inferential statistics.

Keywords: language processing, American Sign Language, visual-world paradigm, inferential statistics

Lieberman, Borovsky, Hatrak, and Mayberry (hereafter LBHM, 2015) applied the visual-world paradigm (Cooper, 1974; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995) to the study of sign processing in American Sign Language (ASL). I argue that whereas LBHM makes an important methodological contribution by showing that the visual-world paradigm can be used to examine the time course of lexical activation and the activation of phonological and semantic competitors during sign processing, their primary theoretical claim—that the mental lexicon of late learners is structured differently from that of native signers—is not supported by either the data pattern or the statistical analyses.

LBHM (2015) assessed phonological competitor effects, which they assumed reflect the activation of sublexical representations, and semantic competitor effects, which they assumed reflect the activation of lexical representations, in a group of native signers and in a group of late-learning signers. (For purposes of this commentary, I assume that both of these assumptions are correct.) They claimed that their results provide evidence that there are structural differences in the organization of the mental lexicon of native signers and late learners. Specifically, they asserted that "only native signers demonstrated early and robust activation of sublexical features of signs during real-time recognition" (p. 1130), that "late-learning signers . . . did not show evidence for real-time activation of sublexical features of sign" (p. 1137), and that "sensitivity to the phonological and semantic relationships among the pictures by the late-learners was evident only later in the time course of lexical recognition (i.e., after the stimulus sign had completely unfolded)" (p. 1137).

Here, I provide a critical evaluation of these claims. First, I argue that none of these claims are supported by LBHM's (2015) results. The data pattern, as reflected in a standard data visualization method for visual-world studies, proportion of fixations as a function of time, suggests that there is neither a quantitative nor a qualitative difference in the time course of phonological and semantic competitor effects between native and late-learning signers. Second, I discuss how three problematic uses of inferential statistics resulted in reaching a set of conclusions that is inconsistent with the data pattern, namely:

1. inappropriate interpretation of differences in statistical significance,
2. insufficient consideration of evidence for low statistical power, and
3. unprincipled analyses involving the use of multiple comparisons.

Third, I provide a reanalysis of LBHM's data that yields no substantial evidence for their claims.

Lieberman et al.'s (2015) Study

Experimental Design, Hypothesis, and Predictions

Participants saw a visual display consisting of a rectangular array of four pictures. After 750 ms, a central fixation cross appeared. As soon as the participant fixated the cross, a video appeared in the center of the screen and started playing. The model in the video produced a sign referring to one of the pictures, and 500 ms later a small cursor appeared in the center of the screen.

The writing of this article was supported by National Institutes of Health Grant HD073890 to Michael K. Tanenhaus. I thank Dale Barr, Florian Jaeger and Michael Tanenhaus for helpful comments and suggestions. I would also like to thank Amy Lieberman for kindly sharing her data.

Correspondence concerning this article should be addressed to Anne Pier Salverda, University of Rochester, Department of Brain and Cognitive Sciences, Meliora Hall, Box 270268, Rochester, NY 14627-0268. E-mail: asalverda@mail.bcs.rochester.edu

Participants used the computer mouse to move the cursor onto the target picture. There were four experimental conditions. In the unrelated condition, the display included three pictures that were phonologically and semantically unrelated to the target sign. In the phonological condition, the sign associated with one picture was phonologically related to the target sign (e.g., it overlapped in handshape but differed in movement of the hand). In the semantic condition, one of the pictured objects was semantically related to the target. In the phono-semantic condition, the display included a phonological and a semantic competitor.

Participants' eye movements were monitored throughout the experiment. Of interest was the time course of fixations to the target and competitor(s) relative to fixations to the unrelated distractor(s). LBHM (2015) hypothesized that the lexicon of late-learning signers, in contrast to that of native signers, does not incorporate sublexical representations. They therefore predicted that real-time phonological competitor effects would be apparent only in native signers, whereas semantic competitor effects would be apparent in both native and late-learning signers.

Results

Rapid, incremental processing of sign language. The results of Experiment 1, with native signers, and Experiment 2, with late-learning signers, are presented in Figures 3 and 5, respectively, in LBHM (2015).[1] These figures present the proportion of fixations to each picture in the visual display over time, relative to the onset of the target word. Fixation proportions over time, which were first introduced by Allopenna, Magnuson, and Tanenhaus (1998), have proved to be a useful data visualization tool that is widely used for looking at visual-world time course data (Barr, 2008; Mirman, Dixon, & Magnuson, 2008). Fixation proportions do not represent all aspects of the eye-movement record (which at the trial level consists of a sequence of saccades, fixations, and blinks). However, in simple word-recognition experiments, where participants typically look at related alternatives before they quickly converge on the target, fixation proportions over time map relatively transparently onto participants' eye-movement behavior (for a discussion of issues concerning the interpretation of linguistically mediated saccades, see Salverda, Brown, & Tanenhaus, 2011). Although apparent differences in fixation proportion curves do not always map onto statistically reliable differences, given the ubiquitous presence of noise in experimental data, there are, to the best of my knowledge, no documented examples of reliable effects in eye-movement data in visual-world studies that are not reflected in the pattern of fixation proportions over time.

The fixation proportion curves presented in LBHM (2015) provide clear evidence for rapid and incremental processing of signs in native and late-learning signers. Fixation proportions to the target started to rise approximately 500 ms after the onset of the sign, well before its offset. (The duration of each sign had been equated across items to 666 ms.) Soon after, fixation proportions to the phonological and semantic competitor increased relative to the distractor for native signers as well as late-learning signers. On the basis of results obtained in visual-world studies with spoken language, a conservative estimate is that fixations occurring within 200 ms after sign offset were programmed during the presentation of the sign (Salverda, Kleinschmidt, & Tanenhaus, 2014). At 866 ms, that is, 200 ms after sign offset, native and late-learning signers were much more likely to fixate the target, phonological competitor or semantic competitor than the distractor(s). These results demonstrate that the sign was processed rapidly and incrementally, as it unfolded in time, and that processing of the sign resulted in the rapid and real-time activation of phonological and semantic competitors (cf. respectively, Allopenna et al., 1998, and Huettig & Altmann, 2005, for results obtained with spoken language).

Competitor effects in native signers versus late-learning signers. The main focus of LBHM's (2015) article is a theoretical question: Might the structure of the mental lexicon of late learners be different from that of native signers? They examined this question by contrasting performance across experimental conditions between native signers, in Experiment 1, and late-learning signers, in Experiment 2. These late learners had been exposed to ASL no sooner than the age of 5 and ranged in their experience with ASL from 5 to 39 years. LBHM hypothesized that whereas the lexicon of native signers is organized at a sublexical level of representation, the lexicon of late-learning signers is not. Consequently, they predicted a phonological competitor effect in native signers but not in late-learning signers. They also hypothesized that the lexicon of both groups of signers is organized similarly at a lexical level of representation and therefore predicted that there would be no difference in semantic competitor effects between the two groups.

Figures 3 (native signers) and 5 (late-learning signers) in LBHM (2015) present the relevant data. Fixation proportions to the target started to rise at approximately 450–500 ms after sign onset for native signers and at about 500–550 ms after sign onset for late-learning signers. Subsequently, fixation proportions to the target increased more rapidly and reached a higher peak for late-learning signers than for native signers. For both groups of participants, fixation proportions to the semantic competitor started to diverge from fixation proportions to the distractor around 550 ms, and fixation proportions to the phonological competitor did so around 650–700 ms after sign onset. Crucially, fixation proportions to the phonological competitor increased more rapidly and reached a higher peak, around 900 ms, for late-learning signers than for native signers. Taking into account a delay of 200 ms for programming an eye movement in response to linguistic input, fixation proportions to the phonological competitor peaked around the offset of the target word (666 + 200 ms). Fixation proportions to the semantic competitor showed a similar pattern. Thus, the pattern of competitor fixation proportions over time suggests a similar time course of phonological and semantic competitor activation for native and late-learning signers, with somewhat stronger phonological and semantic competitor effects for late-learning signers than for native signers. This data pattern is not in line with LBHM's predictions and is inconsistent with their conclusions that "late-learning signers . . . did not show evidence for real-time activation of sublexical features of sign" (p. 1137) and that "sensitivity to the phonological and semantic relationships among the

[1] The interpretation of the results as described in this section is contingent on the proportion of fixation plots as presented in LBHM (2015). Note that these figures present data collapsed across all four experimental conditions—the unrelated, phonological, semantic, and phono-semantic condition.

pictures by the late-learners was evident only later in the time course of lexical recognition (i.e., after the stimulus sign had completely unfolded)" (p. 1137).

Inferential Statistics

LBHM's (2015) conclusions are based on a statistical analysis of their data that suffers from a set of serious issues, which illustrates how incorrect use of inferential statistics can result in conclusions that are inconsistent with the actual data pattern. Examples of similar statistical errors and inaccuracies in interpretation are unfortunately fairly common in the scientific literature. In the article under discussion, the coincidence of a number of errors in the application and interpretation of statistical analyses conspired to generate the impression that there are differences in performance between native and late-learning signers, whereas no such difference is apparent from the data. Next, I discuss each point in detail.

Contrasting Statistically Significant and Statistically Nonsignificant Results

LBHM's (2015) conclusions regarding differences in the activation of phonological and semantic competitors between native and late-learning signers hinge upon a comparison between statistically significant and statistically nonsignificant results. An omnibus analysis of variance (ANOVA) was performed on each of a set of eye-movement measures, with condition as a four-level factor (unrelated, phonological, semantic, phono-semantic). Each of these ANOVAs generated statistically significant results of condition for native signers but not for late-learning signers. For instance, LBHM found a statistically significant effect of condition on the mean latency to generate a saccade to the target for native signers (reported $p < .005$) but not for late-learning signers ($p = .17$). From the results of these two statistical tests, they concluded that the presence of phonological and semantic competitors had a different effect on mean saccade latencies for native signers than for late-learning signers. This conclusion is invalid, because a difference in statistical significance is itself not statistically significant (Gelman & Stern, 2006; see also Nieuwenhuis, Forstmann, & Wagenmakers, 2011). To establish that there is a statistically significant *difference* in the outcome of a measure between native and late-learning signers, a statistical test that incorporates data from both groups is required. For instance, for the analysis just described, a two-way ANOVA with the factors group (native vs. late-learning) and condition would have to reveal a significant interaction between group and condition. Instead, LBHM based their conclusions on differences in statistical significance associated with tests performed separately on data from native signers and data from late-learning listeners. These results offer no sound statistical evidence that any of the measures examined yielded *significantly different results* between native and late-learning signers.

Evidence for Low Statistical Power

Because LBHM (2015) did not perform the critical statistical comparison that is needed to support their main claim, a close evaluation of the summary statistics and the results of their statistical analyses for data from late-learning signers is in order. An important concern that arises is that these data may in fact show qualitatively similar patterns across experimental conditions as found for native signers but that the statistical analyses failed to detect these patterns. This would indicate that their study did not have sufficient statistical power.

The power of a statistical test is the probability that it will lead to rejection of the null hypothesis (Cohen, 1988; for a short introduction see Cohen, 1992). Statistical power depends on the signal-to-noise ratio, the amount of data, and the choice of alpha significance criterion. Given a particular significance criterion ($\alpha < .05$ in the article under discussion), low statistical power can be due to a low signal-to-noise ratio and/or insufficient data. Under conditions of low power, inferences based on whether a statistical test yields a significant result are likely to be misleading: Nonsignificant results do not provide substantive evidence for the absence of an effect in the population, and significant results should be treated with caution because low power increases the likelihood that a statistically significant result is a false positive (Button et al., 2013).

Unfortunately, LBHM (2015) failed to consider the statistical power of their experiments. This is particularly problematic because they predicted a null result in the phonological competitor condition for late-learning signers. Of importance, the late-learning signers were less homogenous, as a group, than were the native signers (due to substantial variation in the age at which they started learning ASL and the number of years of experience with the language). This yields an a priori expectation for a lower signal-to-noise ratio, and thus reduced statistical power, for the late learners. But the only aspect of LBHM's design that compensates for this likely reduction in statistical power to detect a phonological competitor effect is a relatively small increase in sample size from 18 native signers, in Experiment 1, to 21 late-learning signers, in Experiment 2.

It is important to note that LBHM's (2015) analysis of the results for late-learning signers yields a particular pattern of statistically nonsignificant results that suggests that their experiment had low statistical power. For instance, an omnibus ANOVA on mean saccade latencies to the target picture for late-learning signers revealed a statistically nonsignificant effect of condition, with $p = .17$. The authors interpreted this p value only as not significant (given $\alpha = .05$). This categorical assessment is problematic for two reasons. First, the p value is relatively low, an indication that the nonsignificant result may be due to a lack of statistical power. Second, the pattern of mean saccade latencies across conditions for late-learning signers is qualitatively identical to that observed for native signers: The order of conditions as a function of mean saccade latency is the same (see Figure 1). If, as the authors argued, there was no effect of condition on mean saccade latencies for late-learning signers, then the odds that the same qualitative differences in saccade latencies would be found across conditions for late-learning signers as for native signers is 1 in 24 (because there are 4! possible orders of four conditions). Taken together, the relatively low p value of .17 and the qualitative match between the data for native and late-learning signers suggest that the lack of a statistically significant effect of condition on mean saccade

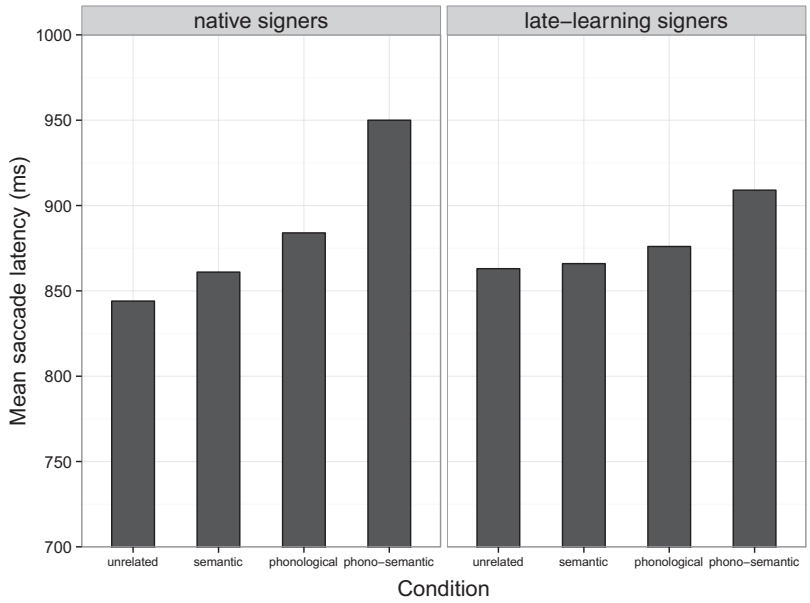

Figure 1. Mean saccade latencies across conditions for native and late-learning signers in Lieberman, Borovsky, Hatrak, and Mayberry (2015). Note that the corresponding Figure 6a in their article does not present the correct mean values for late-learning signers, as reported in their main text (A. M. Lieberman, personal communication, May 21, 2015).

latencies for late-learning signers is due to low statistical power.[2]

Analyses of additional measures obtained from the data of late-learning signers also yielded relatively low p values for the proportion of looks to the target ($p = .18$), the duration of the first fixation to the target picture ($p = .24$), and the total number of fixations on all pictures ($p = .26$). (An analysis on the total number of fixations on the target picture yielded a much higher p value of .93). The authors treated this collection of statistically nonsignificant effects as converging evidence that there was no effect of condition on these measures in late-learning signers. This is a misinterpretation of the logic of null-hypothesis significance testing (see Hauer, 2004, for a comprehensive discussion). In aggregate, the high incidence of relatively low p values across statistical tests on a variety of measures indicates that these measures had low statistical power. Therefore, these statistical analyses do not provide convincing evidence that mean scores on these measures do not vary as a function of condition in the population of late-learning signers, or for the theoretical claim that the lexicon of late-learning signers is structured differently from that of native signers.

Unprincipled Statistical Analyses and Multiple Comparisons

Competitor fixation proportions. LBHM's (2015) research question concerns the activation of phonological and semantic information during the processing of a sign and the degree of activation of such information in native signers and late-learning signers. In their experiments, fixations to the phonological and semantic competitors of a target sign provide the most obvious and direct way to estimate the activation of phonological and semantic information. A large body of research on spoken-word recognition has established that competitor fixations provide a robust and sensitive measure of the activation of phonological (McMurray, Tanenhaus, & Aslin, 2002; Salverda, Dahan, & McQueen, 2003) and semantic information (Huettig, Quinlan, McDonald, & Altmann, 2006). LBHM reported an analysis on the mean proportion of fixations to the competitor(s) and the distractor(s) during a time window from 600 to 1,800 ms following sign onset.[3] An omnibus ANOVA with the factor picture type (phonological competitor and/or semantic competitor, distractor) revealed that there were differences in mean fixation proportions in the phonological con-

[2] Note that all three measures for which LBHM (2015) provided detailed summary statistics across conditions (including the mean proportion of fixations to the target and the video; see Figures 4 and 6 in LBHM) reveal differences between conditions that are qualitatively similar for native versus late-learning signers. This is particularly true for the post hoc comparisons reported by LBHM, which contrast the semantic and unrelated to the phonological and phono-semantic conditions for fixations to the target, and the semantic to the phono-semantic and the phonological conditions for fixations to the video.)

[3] For purposes of this commentary, I assume that these analyses have reasonable validity. However, the use of ANOVA to analyze fixation proportions is problematic for several reasons. First, proportions are bound between 0 and 1 and not normally distributed. Second, differences in proportions are not equivalent across the scale: A .01 change in proportions from .01 to .02 (a 100% increase) is more substantial than a change from .10 to .11 (a 10% increase). In order to analyze these data within an ANOVA framework, a log-odds transformation should be applied (see Jaeger, 2008). An additional problem arises in the analysis under discussion: Fixation proportions to pictures within the same display are not independent. (Note that these dependencies can be resolved by transforming fixation proportions to two or more pictures into a ratio. For instance, if A is fixated more than is B, the ratio $\frac{A}{A+B}$ exceeds .5.)

dition for late-learning signers and in the semantic condition as well as the phono-semantic condition for late-learning and native signers.[4] LBHM's analyses of competitor fixations thus provide evidence for phonological competitor effects in late-learning signers and semantic competitor effects in native and late-learning signers. These results support the conclusion that native and late-learning signers show similar semantic competitor effects, as LBHM predicted. However, they also suggest that, if there are any differences in phonological effects, late-learning signers actually show *stronger* phonological competitor effects than do native signers. This is contrary to LBHM's prediction that native but not late-learning signers would show such an effect.

Additional eye-movement measures. LBHM (2015) examined six additional eye-movement measures to further evaluate phonological and semantic competitor effects:

1. the latency of the first saccade to the target,
2. the proportion of fixations to the target,
3. the proportion of fixations to the video,
4. the duration of the first fixation to the target,
5. the number of fixations to the target, and
6. the total number of fixations to all pictures.

Statistical analyses of the results from the additional measures were not performed in a principled way and did not yield a comprehensive and robust pattern of results. This suggests that these measures had low power (see also the Evidence for Low Statistical Power section and the Reanalysis of the Data section) and possibly low validity. Each measure was analyzed with an omnibus ANOVA including data from all four experimental conditions. However, an omnibus ANOVA that yields a significant result does not provide information specifying which conditions differ from each other (Abelson, 1995, p. 105, justly compared omnibus testing to "playing the guitar with mittens on"). An omnibus ANOVA needs to be complemented by principled (and a priori determined) follow-up tests that address predictions derived from the research hypothesis by examining differences between specific conditions. For instance, one set of follow-up tests that is pertinent to LBHM's (2015) hypothesis about the activation of phonological competitors would contrast the phonological condition to the unrelated condition (cf. the analysis of competitor fixation proportions as discussed in the Competitor Fixation Proportions section; see also the Reanalysis of the Data section).

Omnibus ANOVAs on data from native signers yielded significant effects of condition for each of the additional measures. No follow-up tests were reported for three of these measures (the duration of the first fixation to the target, the number of fixations to the target, and the total number of fixations to all pictures), suggesting that they did not yield statistically significant differences between conditions that could be used to evaluate the research hypotheses. For the remaining three additional measures, one or two follow-up tests were reported. However, these tests were not selected in a principled way. For instance, none of these tests contrasted a particular competitor condition with the unrelated condition. Moreover, the follow-up tests did not yield robust patterns of results that replicated across measures, which further questions their validity in the context of the research hypothesis.

Table 1
ANOVA Test Statistic and p Values for the Group × Condition Interaction for the Six Additional Measures Reported in LBHM

Measure	$F(3, 111)$	p
Latency of first saccade to target	1.07	.36
Proportion of fixations to target	1.58	.20
Proportion of fixations to video	2.36	.075
Duration of first fixation to target	1.34	.27
Number of fixations to target	1.25	.30
Total number of fixations	1.10	.35

Note. ANOVA = analysis of variance; LBHM = Lieberman, Borovsky, Hatrak, and Mayberry (2015).

The fact that the follow-up tests reported were not appropriately motivated and that different follow-up tests were reported for each measure raises the concern that these tests, which were reported as "planned comparisons" (suggesting that the particular conditions contrasted in each test had been determined ahead of time), are more appropriately examples of post hoc follow-up tests. That is, the conditions contrasted in these tests were determined at least in part by the pattern of results observed across conditions for a particular measure. Alternatively, many possible planned comparisons were deemed compatible with aspects of the research hypothesis, but only a small subset of those comparisons yielded significant results. In either case, the analysis procedure involved performing multiple comparisons in order to identify which particular follow-up tests yielded statistically significant results. Such a procedure increases the likelihood of obtaining false positives, and appropriate corrections need to be applied to the test results. It is unclear whether the four follow-up tests reported by LBHM (2015)—three of which had a reported p value of $< .05$, and one of which was reported only as statistically significant—would yield statistically significant results if corrections for multiple comparisons had been applied.

LBHM's (2015) conclusion that there are differences in the organization of the mental lexicon between native and late-learning signers rests on the assumption that results obtained for the additional measures were more meaningful and more insightful than were those obtained for fixation proportions to the phonological and semantic competitors, which yielded results that contradict LBHM's conclusion. However, the fact that the statistical analyses of the additional measures did not yield robust and comprehensive results (with respect to LBHM's predictions) indicates that they likely have insufficient statistical power and low validity. Consequently, results obtained with these measures have little or no evidential value with respect to the research hypothesis. In contrast, results obtained in the analysis of competitor fixation proportions were clearer and more comprehensive, but these results provide no evidence for LBHM's hypothesis that the lexicon of

[4] Although mean fixation proportions for each type of picture and the proportion of fixation plots in LBHM (2015) suggest that these significant effects were likely due to differences in fixation proportions to the competitor(s) versus the distractor(s), follow-up tests should have assessed these contrasts directly (see also the Additional Eye-Movement Measures section).

Table 2
Effect Sizes Contrasting the Phonological and Unrelated Conditions for the Six Additional Measures Reported in LBHM

Measure	Native signers			Late-learning signers		
	Unrelated	Phonological	Effect size	Unrelated	Phonological	Effect size
Latency of first saccade to target (ms)	844	884	40	863	876	13
Proportion of fixations to target	.51	.44	−.07	.54	.51	−.03
Proportion of fixations to video	.42	.45	.04	.36	.38	.03
Duration of first fixation to target (ms)	562	535	−27	529	517	−12
Number of fixations to target	1.55	1.43	−.12	1.81	1.84	.02
Total number of fixations	2.89	3.06	.17	3.60	3.68	.08

Note. Note that across measures, with the exception of the number of fixations to the target, the effect size is in the same direction for native and late-learning signers. LBHM = Lieberman, Borovsky, Hatrak, and Mayberry (2015).

late-learning signers is organized differently from that of native signers.

Reanalysis of the Data

In order to provide additional support for my claim that LBHM's (2015) data do not provide substantial evidence for qualitatively different phonological competitor effects between native and late-learning signers, I reanalyzed their data. I used the same analysis as adopted by the authors, ANOVA, and examined performance across all six additional measures that LBHM presented as supporting their claim, using the same data that they analyzed: the mean value for each measure as a function of condition, for each participant.

Omnibus ANOVA

A repeated-measures omnibus ANOVA with the factors group (native vs. late-learning) and condition (unrelated, phonological, semantic, phono-semantic) revealed that the Group × Condition interaction was not statistically significant for any of the six measures (see Table 1). The data are therefore insufficient to conclude that there are differences in the activation of phonological and semantic competitors between these two groups of signers, contrary to LBHM's (2015) main conclusion.

In my evaluation of the nonsignificant Group × Condition interaction, two additional things are worth noting. First, although each of the six tests reported in Table 1 is associated with a Type I error rate of .05, the familywise Type I error rate (across the set of six measures) is substantially higher. Appropriate corrections for multiple comparisons would therefore result in an alpha significance level substantially below .05. Second, it is important to keep in mind that the mere presence of a Group × Condition interaction (if it had been found) could reflect a qualitative or quantitative difference in competitor activation between native and late-learning signers. In order to support LBHM's (2015) claims, one would need to supplement a significant Group × Condition interaction by follow-up tests showing that the interaction reflects qualitative differences in the activation of phonological competitors between native and late-learning signers—for example, the activation of phonological competitors for native but not for late-learning signers—as opposed to a difference in the *degree* of activation.

Phonological Competitor Effect

LBHM's (2015) hypothesis that the lexicon of late-learning signers, in contrast to that of native signers, does not incorporate sublexical representations leads to the prediction that phonological competitor effects should be observed for native but not for late-learning signers. I evaluated the presence (and degree) of phonological competition for native versus late-learning signers by contrasting performance in the phonological versus unrelated condition. Notably, LBHM did not systematically evaluate or report this contrast.

I first computed the effect sizes, that is, the difference between the phonological and unrelated condition across the set of six measures (see Table 2). For native signers, effect sizes were consistent with a phonological competitor effect for all six measures (e.g., a longer mean saccade latency to the target in the phonological condition than in the unrelated condition). However, for late-learning signers, effect sizes were also consistent with a phonological competitor effect for five of the six measures. If there were no phonological competitor effect for late-learning signers, then the odds that this particular pattern, or one more consistent with the direction of the effect sizes for native signers, would arise by chance is 7/64, that is, approximately 1 in 9. Therefore, the pattern of effect sizes suggests that there was at best a difference in the *degree* of phonological competition between native and late-learning signers, contradicting LBHM's (2015) claim that there was a qualitative difference (i.e., a phonological competitor effect for native but not for late-learning signers).

I then reanalyzed the data for the six additional measures, contrasting performance in the phonological condition with the unrelated (baseline) condition (see Table 3). A Group (native vs. late-learning signers) × Condition (unrelated, phonological, semantic, phono-semantic) ANOVA[5] revealed that the interaction between group and the phonological versus unrelated contrast was not statistically significant for any of the six measures (see Table 3). The data are thus insufficient to conclude that the phonological competitor effect for native signers is different from that for late-learning signers. For native signers, the simple effect of the

[5] An ANOVA on the subset of data from the unrelated and phonological conditions yielded a qualitatively similar pattern of results and led to the same conclusions as discussed for the ANOVA including data from all four conditions.

Table 3
ANOVA Test Statistic and p Values, Contrasting the Phonological and Unrelated Conditions for the Six Additional Measures Reported in LBHM

	Native vs. late-learning signers		Simple effect: Native signers		Simple effect: Late-learning signers	
Measure	$F(1, 111)$	p	$F(1, 111)$	p	$F(1, 111)$	p
Latency of first saccade to target	0.57	.45	2.38	.13	0.31	.58
Proportion of fixations to target	1.83	.18	9.19	.0030	1.64	.20
Proportion of fixations to video	0.23	.64	4.10	.045	2.21	.14
Duration of first fixation to target	0.15	.70	0.84	.36	0.18	.67
Number of fixations to target	1.47	.23	1.91	.17	0.09	.77
Total number of fixations	0.37	.54	2.64	.11	0.74	.39

Note. Note that the Group × Contrast interaction is not statistically significant for any of the measures. ANOVA = analysis of variance; LBHM = Lieberman, Borovsky, Hatrak, and Mayberry (2015).

phonological versus unrelated contrast was statistically significant for just two of the six measures (proportion of fixations to the target; proportion of fixations to the video). This suggests that the additional measures, with the exception of the proportion of fixations to the target, had low statistical power to detect a phonological competitor effect for the native signers. For late-learning signers, the simple effect of the phonological versus unrelated contrast was not statistically significant for any of the measures (but note that the lowest p values were obtained for the two measures that yielded a significant effect for native signers). It is important to note that this pattern of nonsignificance is not particularly informative, given that the analysis of data from native signers indicated that these measures had low statistical power to detect a phonological competitor effect. Moreover, it is reasonable to have an a priori expectation that late-learning signers would show weaker phonological competitor effects than would native signers, because they are not as skilled in processing sign language. Other things being equal, statistical power decreases as effect size decreases, thus making it likely that LBHM's (2015) experiment with late-learning signers did not have sufficient statistical power to detect a phonological competitor effect that was weaker than that for native signers.

Conclusions

A critical examination of results reported in LBHM (2015) reveals no support for their hypothesis that the mental lexicon of late-learning signers is organized differently from that of native signers. Their conclusion that late-learning signers, in contrast to native signers, do not activate sublexical representations during the processing of a sign is not warranted by the data and is not supported by principled statistical analyses of those data.

In recent years, there has been increased concern that inappropriate use of inferential statistics in the behavioral science literature results in researchers' drawing conclusions that are not strongly supported by the data. Much of the concern has focused on the fact that results might not replicate, therefore slowing progress in the field. Now that people are generally more aware of concerns about power and failures to replicate receive more attention in the literature, it might be tempting to overlook what might appear to be departures from best practice in statistical analyses as not particularly damaging, especially when the overall conclusion seems reasonable.

A critical analysis of LBHM (2015) is revealing in that it shows how departures from best practice statistical analyses can lead to conclusions that not only do not receive strong support from the statistical analyses but that are also inconsistent with the actual data pattern. This became clear only because, to their credit, LBHM combined their use of inferential statistics with data visualization that showed the important patterns in the data. This highlights the importance of both following best practice in statistics and using data visualization techniques to verify that the conclusions from inferential statistics have face validity. When those conclusions do not map onto the data, it is critically important to reevaluate the data (which may be insufficient), the measures (which may lack validity), and the statistical analyses (which may be inappropriate or insufficient).

To conclude, LBHM (2015) makes a clear contribution to the ASL-processing literature by demonstrating that the visual-world paradigm can be used to study the real-time processing of ASL. Proficient signers, like users of spoken languages, process linguistic input rapidly and incrementally. Moreover, both semantically and phonologically related pictures yielded clear competitor effects, which should prove useful in future studies. However, the claim that the lexicon of late-learning signers is organized differently from that of native signers receives no substantive support from the results of their study or the statistical analyses of those results. Rather, LBHM provided a case study in how departures from best practice in the use of inferential statistics can lead to conclusions that are incompatible with the data.

References

Abelson, R. P. (1995). *Statistics as principled argument*. Hillsdale, NJ: Erlbaum.

Allopenna, P., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language, 38,* 419–439. http://dx.doi.org/10.1006/jmla.1997.2558

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

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14,* 365–376. http://dx.doi.org/10.1038/nrn3475

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159. http://dx.doi.org/10.1037/0033-2909.112.1.155

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology, 6,* 84–107. http://dx.doi.org/10.1016/0010-0285(74)90005-X

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *American Statistician, 60,* 328–331. http://dx.doi.org/10.1198/000313006X152649

Hauer, E. (2004). The harm done by tests of significance. *Accident Analysis & Prevention, 36,* 495–500. http://dx.doi.org/10.1016/S0001-4575(03)00036-8

Huettig, F., & Altmann, G. T. M. (2005). Word meaning and the control of eye fixation: Semantic competitor effects and the visual world paradigm. *Cognition, 96*(1), B23–B32. http://dx.doi.org/10.1016/j.cognition.2004.10.003

Huettig, F., Quinlan, P. T., McDonald, S. A., & Altmann, G. T. M. (2006). Models of high-dimensional semantic space predict language-mediated eye movements in the visual world. *Acta Psychologica, 121,* 65–80. http://dx.doi.org/10.1016/j.actpsy.2005.06.002

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59,* 434–446. http://dx.doi.org/10.1016/j.jml.2007.11.007

Lieberman, A. M., Borovsky, A., Hatrak, M., & Mayberry, R. I. (2015). Real-time processing of ASL signs: Delayed first language acquisition affects organization of the mental lexicon. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41,* 1130–1139. http://dx.doi.org/10.1037/xlm0000088

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition, 86*(2), B33–B42. http://dx.doi.org/10.1016/S0010-0277(02)00157-9

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language, 59,* 475–494. http://dx.doi.org/10.1016/j.jml.2007.11.006

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature Neuroscience, 14,* 1105–1107. http://dx.doi.org/10.1038/nn.2886

Salverda, A. P., Brown, M., & Tanenhaus, M. K. (2011). A goal-based perspective on eye movements in visual world studies. *Acta Psychologica, 137,* 172–180. http://dx.doi.org/10.1016/j.actpsy.2010.09.010

Salverda, A. P., Dahan, D., & McQueen, J. M. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition, 90,* 51–89. http://dx.doi.org/10.1016/S0010-0277(03)00139-2

Salverda, A. P., Kleinschmidt, D., & Tanenhaus, M. K. (2014). Immediate effects of anticipatory coarticulation in spoken-word recognition. *Journal of Memory and Language, 71,* 145–163.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995, June 16). Integration of visual and linguistic information in spoken language comprehension. *Science, 268,* 1632–1634. http://dx.doi.org/10.1126/science.7777863

Received July 10, 2015
Revision received January 20, 2016
Accepted January 22, 2016 ■