UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Hartshorne, J. K., et al. (2019). The Meta-Science of Adult Statistical Word Segmentation: Part 1. *Collabra: Psychology,* 5(1): 1. DOI: https://doi.org/10.1525/collabra.181

ORIGINAL RESEARCH REPORT

The Meta-Science of Adult Statistical Word Segmentation: Part 1

Joshua K. Hartshorne, Lauren Skorb, Sven L. Dietz, Caitlin R. Garcia, Gina L. Iozzo, Katie E. Lamirato, James R. Ledoux, Jesse Mu, Kara N. Murdock, Jon Ravid, Alyssa A. Savery, James E. Spizzirro, Kelsey A. Trimm, Kendall D. van Horne and Juliani Vidal

We report the first set of results in a multi-year project to assess the robustness – and the factors promoting robustness – of the adult statistical word segmentation literature. This includes eight total experiments replicating six different experiments. The purpose of these replications is to assess the reproducibility of reported experiments, examine the replicability of their results, and provide more accurate effect size estimates. Reproducibility was mixed, with several papers either lacking crucial details or containing errors in the description of method, making it difficult to ascertain what was done. Replicability was also mixed: although in every instance we confirmed above-chance statistical word segmentation, many theoretically important moderations of that learning failed to replicate. Moreover, learning success was generally much lower than in the original studies. In the General Discussion, we consider whether these differences are due to differences in subject populations, low power in the original studies, or some combination of these and other factors. We also consider whether these findings are likely to generalize to the broader statistical word segmentation literature.

Keywords: language acquisition; word segmentation; statistical learning; replication

Science is the process of becoming less wrong. Meta-science is the study of how to become less wrong more quickly. All fields engage in some amount of meta-science, but it is particularly appropriate for psychologists. Psychology is the study of human behavior, and science is an especially interesting and complex human behavior.

The efficiency of the current scientific paradigm has lately occasioned a great deal of debate, particularly within psychology. At issue are false positive rates, false negative rates, the focus on theoretical ground-breaking vs. precise measurement, the efficacy of the peer review system, academic incentive systems, the proper role of meta-analysis, and cross-cultural generalizability, among others (Anderson et al., 2016; Asendorpf et al., 2013; Bakker et al., 2012; Button et al., 2013; Crane, 2017; Ferreira & Henderson, 2017; Gibson & Fedorenko, 2010, 2011; Gilbert et al., 2016; Gray & Wegner, 2013; Hartshorne & Schachner, 2012; Henrich et al., 2010; Ioannidis, 2005, 2012; Judd et al., 2012; Kerr, 1998; Maxwell et al., 2015; Nosek & Lakens, 2014; Nosek et al., 2012; Open Science Collaboration, 2015; Pashler & Harris, 2012; Simons et al., 2014; Stroebe, 2016; Stroebe et al., 2012; Vul et al., 2009; Zwaan et al., n.d.). While none of these debates are new or unique to recent years, recently there is an increased interest in informing these debates through empirical studies: that is, meta-science (Aarts & LeBel, 2016; Bakker et al., 2016; Button et al., 2013; Camerer et al., 2016; Cova et al., 2018; Ebersole et al., 2016; Fraley & Vazire, 2014; Frank et al., 2017; Klein et al., 2014; Mahowald et al., 2017; Makel et al., 2012; Open Science Collaboration, 2015; Schweinsberg et al., 2016; Shen et al., 2011; Vankov et al., 2014). This empirical work has put in-principle arguments about efficiency on a firmer footing.

Nonetheless, data remain sparse. Systematic studies of replicability, sample sizes, and effect sizes are concentrated in certain parts of the field, particularly social psychology, clinical psychology, and fMRI research (Button et al., 2013; Chase & Chase, 1976; Fraley & Vazire, 2014; Mone et al., 1996; Richard et al., 2003; Shen et al., 2011; Vankov et al., 2014). Even within those subfields, the data are patchy: The largest database of replications in psychology (curatescience.org) lists only 1,058 (for details on the project, see Aarts & LeBel, 2016). The sketchiness of the data leaves considerable room for debate about the state of the field as a whole (cf. Bakker et al., 2012; Klatzky et al., 2018).

This paper reports the first 8 replications in a long-term project to replicate as many as possible of the

Department of Psychology, Boston College, US

Corresponding author: Joshua K. Hartshorne (joshua.hartshorne@bc.edu)

100+ experiments in the adult statistical word segmentation literature.[1] This narrow focus is a departure from most replication and meta-science studies, which tend to focus on entire subdisciplines if not the field as a whole.

Our more narrow focus has a straightforward explanation: the statistical word segmentation literature is of monumental theoretical importance. The finding that humans can learn a great deal about language without feedback and without overt tasks suggests at least a partial solution to one of the overriding concerns of modern cognitive science: the poverty of the stimulus problem (Ambridge et al., 2009; Baker, 1979; Bowerman, 1988; Brown & Hanlon, 1970; Chemla et al., 2009; Marcus, 1993; Marcus & Berent, 2003; Perfors et al., 2010; Pinker, 1984; Quine, 1960; Romberg & Saffran, 2010; Tomasello, 2009).

However, the simple demonstration that humans can engage in statistical learning is only the first step. In fact, there are myriad learning algorithms for statistical learning (Frank et al., 2010; Goldwater et al., 2009; Kurumada et al., 2013; Pearl et al., 2010; Thiessen, 2017). These include algorithms that instantiate explicit tabulation of transition probabilities, clustering algorithms, memory compression, recurrent neural networks, and inference over generative models—each of which can be instantiated in a variety of manners. Thus, dozens of experiments have gone beyond simply demonstrating that humans can engage in some kind of statistical learning of language, attempting to disentangle the underlying mechanisms. The answer has broad implications: In addition to providing more or less powerful avenues for learning, different statistical learning mechanisms would suggest linguistic representations ranging from highly symbolic to constructionist, and learning theories ranging from strongly empiricist to strongly nativist.

Of statistical learning problems, statistical word segmentation is relatively simple and has been particularly well-explored, providing rich grist for theory. However, given the current epistemic uncertainty plaguing psychology and other fields, it would be ideal to know which findings from this crucial empirical literature can be relied upon. Moreover, to the extent that replicability varies across the literature, we hope to learn what methods tend to produce more robust results, thereby aiding researchers in making discoveries going forward. Finally, we aim to obtain more accurate effect size estimates across this literature, which is increasingly important for developing and testing more precise theories (Frank et al., 2016; Lewis et al., 2016).

Though our focus is on adult statistical word segmentation, we turn to the implications our findings may have for the larger debate about replicability in the General Discussion.

Overview of the Project

There are many empirical questions relevant to the meta-science of psychology: How common are replications (Hartshorne & Schachner, 2012; Makel et al., 2012)? What is the size of the file drawer (Rosenthal, 1979; Spellman, 2012)? What is the prevalence of misanalysis (Hardwicke et al., 2018)? What are typical effect sizes, sample sizes, and power levels (Bezeau & Graves, 2001; Button et al., 2013; Fraley & Vazire, 2014; Mone et al., 1996; Rossi, 1990; Vankov et al., 2014)? How common is data falsification (Fanelli, 2009; Rhoades, 2004)? How reliable is peer review (Mahoney, 1977; Newton, 2010; Olson et al., 2002)? How much do observed effects vary from lab to lab (Frank et al., 2017; Klein et al., 2014)? How does the undergraduate subject population vary across the semester (Ebersole et al., 2016)?

We focus on the following: How reproducible are published studies of adult statistical word segmentation, and how reliable are their reported results? By *reproducible*, we mean that a study can be recreated using the public record. By *reliable*, we mean that claims of the form *under condition X, results Y reliably obtain* hold up when the experiment is faithfully reproduced.

Note that this means we cannot address whether the theoretical conclusions are true. Such a study would likely involve designing new experiments that improve on the methods of published studies. While incredibly useful science, such a study would not address our meta-scientific question, which is how much researchers can rely on the methods and results sections of the papers they read, and under what circumstances. Our question requires not new experiments, but faithful replications, regardless of whether the original study was well conceived.

For similar reasons, we focus on the original methods and results *as reported*. The importance of assessing the reliability of the results as reported in the literature stems from the fact that it is these reports that are evaluated by reviewers, interpreted by readers, and explained (or not) by theorists. If those reports are incorrect, readers are being misinformed and theories are being improperly evaluated. Knowing whether or not this is the case is of fundamental importance to the advancement of science. It is of course also meta-scientifically useful to know how often reports in the literature are unreliable because the method was incompletely or incorrectly described. Answering that question, however, requires a very different method from answering our question. At the very least, one would need to replicate every experiment twice: once as reported and once as corrected by the original authors. This is beyond the scope of the present project. Our approach also means that we did not discuss the replications with the original authors, with one exception detailed below.

Note also that there is no agreed-upon definition for *replication* as it applies to the cognitive and behavioral sciences. This stems from the fact that it is impossible even in theory to exactly reproduce the subjects and cultural context of the original, and often many other methodological details can be reproduced only approximately (like the physical apparata). As a first approximation, something is a replication if the differences in methods are benign enough that any difference in results is more likely to be explained by a false positive (or negative) in the original. Since this is not (currently) a question than can be answered objectively, there are few questions in psychology that occasion as much debate

as whether a particular experiment is a "true" replication of some prior experiment. Even where the replicators, original authors, and two sets of reviewers agree that something is a replication, it is not unprecedented for others to disagree (e.g., Gilbert et al., 2016).

Thus, while we have endeavored wheverever possible to match the original methods in both detail and spirit, there are some differences. Many reflect the passage of time: we use MP3s rather than cassette tapes and keyboard responses rather than paper and pencil. We do not familiarize our subjects with how to use a computer, as some of the experiments from the mid-1990s did. Others reflect choices ancillary to testing itself: we compensate participants monetarily, whereas some of the originals used course credit. In a few places, we have standardized minor aspects of method in order to improve comparability across experiments. For instance, we use a standard sample of 50 subjects per condition (far more than in any of the originals), and all experiments are computerized and programmed using jsPsych (De Leeuw, 2015).[2] The combination of large samples and computerization allows us to randomize the order of test trials separately for each participant. We believe most of these differences are *a priori* unlikely to matter – and if they did, they would prompt a significant revision in our understanding of the original results, and perhaps of the entire field (cf. Perone, 2018). Either way, the differences in results would be of significant theoretical importance. We have exhaustively documented all such methodological decisions in individual preprints for each experiment, allowing the reader to judge for themselves and design follow-up experiments if needed (Garcia, Iozzo, et al., 2017; Garcia, van Horne, & Hartshorne, 2017; Hartshorne, 2017; Hartshorne & Skorb, 2018; Iozzo et al., 2017; Mu et al., 2017; Murdock et al., 2017a, b). Note that our materials, data, and code are available as well (see preprints for details).

There are several methodological issues that deserve more extensive discussion: the statistical analyses, subject population, and speech synthesizer.

Statistical analysis. In many cases, the originally-reported statistical analyses contain significant analytic or conceptual errors. The most common include the treatment of items as fixed effects and the application of linear models (e.g., t-tests and ANOVAs) to non-linear data (e.g., percentage correct) (Baayen et al., 2008; Clark, 1973; Jaeger, 2008). Both of these errors can lead to false positives; the latter can result in false negatives as well.

Thus, in cases where the original analyses contained errors, we analyzed the experiment twice: once according to the original method and once using more appropriate statistical models (e.g., mixed effects logistic regression). For simplicity, we report the revised analyses below and relegate the original, unreliable analyses to the preprints. *However, wherever the different methods of analyses result in different patterns of significance, this is noted in the main text.*

Subject population. In designing our project, we planned on recruiting subjects through Amazon Mechanical Turk (AMT) and screening out inattentive subjects. As far as study design goes, both choices are well-supported by the meta-scientific literature. The quality of AMT data has been confirmed in numerous high-power studies (Berinsky et al., 2012; Buhrmester et al., 2011; Chandler et al., 2014; Crump et al., 2013; Mason & Suri, 2011; Rand, 2012; Rouse, 2015; Shapiro et al., 2013), and using AMT made it possible to test enough subjects to obtain adequate statistical power. Screening out inattentive subjects is best-practice for any study, whether run in the laboratory or online.

However, both choices raise potential concerns for replication. While AMT data is high-quality, the subject population tends to be more diverse in age, ethnicity, SES, and education than are the university-based samples used in the majority of prior studies (Henrich et al., 2010; Ipeirotis, 2010; Paolacci et al., 2010). There is an implicit assumption in the original papers that their findings generalize beyond high-SES psychology majors; otherwise, the findings would have little relevance for typical first-language acquisition. However, this is an untested hypothesis, and we could conceivably obtain different results from our different population. Thus, we directly test the equivalence of the populations in Exp. 1.

Similarly, none of the original studies screened out inattentive subjects. In principle, it would be reasonable to similarly ommit catch trials from the replications: The design of the studies is such that inattentive subjects are unlikely to cause any of the observed effects, and inattentive subjects are not necessarily more likely on AMT than in the lab (Hauser & Schwarz, 2016; J. A. Johnson, 2005). Nonetheless, we felt in this case it would be better to follow best practice than the original method. We test the reasonableness of this decision in Exp. 1.

Speech synthesizer. Studies in this literature frequently use synthesized stimuli. In many cases, the synthesizer used no longer exists. We tried several alternatives. Based on the results of Exp. 1 (below), we elected to default to MBROLA synthesizer (Dutoit et al., 1996), which has been widely used in statistical learning studies (Bonatti et al., 2005; Fernandes et al., 2011; Franco et al., 2011; Frank et al., 2010; Hoch et al., 2013; E. K. Johnson & Tyler, 2010; Kovacs & Mehler, 2009; Perruchet & Desaulty, 2008; Toro et al., 2005).

Description of Experiments

First, we present three replications of Saffran, Newport, & Aslin (1996), Exp. 1, upon which many of the subsequent studies in the literature are based. Across these replications, we vary the population (Amazon Mechanical Turk vs. Boston College undergraduates), the use of attentional screens, and the speech synthesizer. This allows us to investigate the role of these methodological choices, and informs our decision to use Amazon Mechanical Turk, attention screens, and MBROLA. We follow with replications of 5 more experiments chosen from among the more highly-cited papers in the literature: Exps. 1 and 2 from (Saffran et al., 1999), Exps. 1 and 3 from (Finn & Hudson Kam, 2008), and Exp. 1 from (Frank et al., 2010). Because we intend to ultimately replicate all such experiments, no attempt was made to choose a representative set for this first foray.

In every case, we aimed to test 50 subjects per condition. Subjects excluded for inattention were not replaced due to cost considerations.

In our description below, we used a simple metric for replication: was the pattern of significance the same? There are many other more advanced and arguably more informative metrics. Unfortunately, they tend to require the original experiment to be high-powered, which was generally not the case here (Morey & Lakens, 2016).[3] However, since the data are public (see below), other researchers may explore other metrics, including advanced metrics that are not yet developed.

We adopted a number of mechanisms to ensure quality control. Every aspect of data-collection and analysis was conducted programmatically with version-controlled code. This not only eliminates any possibility of experimenter bias or contamination, but also facilitates double-checking every aspect of the workflow. Stimuli and code for each experiment was vetted by JKH. Analysis code and write-ups were vetted by JKH and LS, with additional spot-checking by Miguel Mejia and Hayley Greenough.

For each experiment, we describe all stimuli and measures, exclusions, and analyses. *Exhaustive* descriptions of each experiment are available in individual preprints posted on PsiArXiv (Garcia, Iozzo, et al., 2017; Garcia, van Horne, & Hartshorne, 2017; Hartshorne, 2017; Hartshorne & Skorb, 2018; Iozzo et al., 2017; Mu et al., 2017; Murdock et al., 2017a, b). These write-ups include additional methodological and statistical details and provide links to Open Science Framework repositories that contain all data, materials, and code.

All research was approved by the Boston College Office for Research Protections.

Investigation of the method: Three replications of Saffran, Newport, & Aslin (1996) Exp. 1

We conducted three replications of Saffran, Newport, & Aslin (1996), varying **subject population** (AMT vs. university subject pool), the use of an **attention screen**, and the **choice of speech synthesizer**. We chose this experiment as our starting point because it was this paper, along with a companion paper published the same year, that launched the modern statistical word segmentation literature.

The subject population and attention screens have been discussed in detail above. With regards to the speech synthesizer, Saffran and colleagues generated stimuli using MacinTalk, a commercial speech synthesizer bundled with Apple computers in the 1980s and 1990s. They do not specify which version, and we were in any case unable to obtain working copies of any version. In the intervening years, synthetic speech technology has advanced greatly. However, these same innovations have made them incapable of producing the monotone, constant-rate speech that is critical for many word segmentation studies. The best option we found for approximating the characteristics of the original stimuli was MBROLA, a widely-used open-source speech synthesizer (Dutoit et al., 1996). However, an anonymous reviewer suggested that MBROLA's speech quality is poor, potentially diminishing subjects' ability to learn. Thus, we conducted an additional replication using a modern speech synthesizer: IBM Watson's Text to Speech (IBM, 2017). While this speech synthesizer is unable to produce stimuli as described in the original paper, it does sound significantly more realistic. (To access our stimuli, follow the links provided in the individual experiment preprints.)

Overview of original experiment

Along with Saffran, Aslin, & Newport (1996), Saffran, Newport, & Aslin (1996) (henceforth: SNA96-1) was the paradigm-defining report of statistical word segmentation. There were several key findings. After listening to a novel language made of three-syllable words, adults were able to discriminate words in the language from foils. This was assumed to be due to recognizing the fact that the between-syllable transition probabilities during training were higher within-word than between-word. Evidence for this came from the finding that subjects were better able to recognize words that had higher internal transitional probabilities relative to those with lower internal transitional probabilities. The authors also reported a primacy effect, where subjects were more likely to reject foils that matched the end of a trained word than ones that matched the beginning of a trained word.

Method

The preprints describing each of the three replications in detail can be found at psyarxiv.com/m39yw, psyarxiv.com/qsyd2, and psyarxiv.com/e5c64.

Exp. 1a: AMT+MBROLA

Subjects. 100 individuals were recruited through Amazon Mechanical Turk and participated in exchange for monetary compensation, 98 of whom were native English speakers. These included 50 subjects in the part-word condition (Ages 20–58, M = 32) and 50 subjects in the nonword condition (Ages 20–60, M = 33). 34 subjects in the nonword condition (Ages 20–60, M = 32) and 19 in the part-word condition (Ages 24–58, M = 35) answered all the catch trials correctly.

Materials. The language consisted of four consonants (/p/, /t/, /b/, /d/) and three vowels (/a/, /i/, /u/) which, when combined, rendered an inventory of 12 CV syllables. 11 of these syllables (excluding /di/) were combined to create six trisyllabic words: *babupu, bupada, dutaba, patubi, pidabu,* and *tutibu*. 252 tokens of each of the 6 words were concatenated in a random order, with the stipulation that the same word never occurred twice in a row. Average transitional probability within a word ranged from 43% to 100%. Transitional probabilities between phonemes spanning words ranged from 5.0% to 59%.

These words were then vocalized using MBROLA's us1 voice (Dutoit et al., 1996). Each syllable was produced in context with full coarticulation between syllables. For each syllable, the consonant lasted for 100 ms and the vowel for 177 ms, for a rate of 216 syllables/minute. The resulting sound file was converted from WAV to MP3 using Sound eXchange (SoX) (Bagwell & Contributors, 2015).

For the test phase, six nonword foils and six part-word foils were created. The nonwords (*pibuda, badapu, tapubi, bubita, dubiti, bubibi*) consisted of syllables from the language's syllable inventory which never followed each other in the speech stream, even across word boundaries. Thus, the transitional probabilities between each of the syllables in the nonwords were zero. The part-words either contained the first two syllables of words plus an additional syllable (*pidata, bupabi, babuda*) or contained contained the final two syllables of words, starting with an additional syllable (*bitaba, datubi, titibu*). The original authors explicitly name three of the part-words they used (*pidata, bitaba, bupabi*), and thus we included them. All other foils were randomly generated by a computer program, using the definitions given in the original for nonwords and part-words, and excluding the syllable (/di/) that did not appear in the training. In addition, we generated 6 catch trials, which pitted each word against a foil that contained multiple phonemes that did not appear in the training. Targets and foils were produced by MBROLA in isolation.

Procedure. Subjects were instructed to listen to a 'nonsense' language. They were told that the language contained words, but no meanings or grammar. They were informed that their task was to figure out where the words began and ended. Subjects were given no information about the length or structure of the words or how many words the language contained. They were informed that the listening phase of the experiment consisted of three short blocks, followed by a test of their knowledge of the words in the language. The training was broken into three 7-min listening blocks. Subjects were allowed to take a self-paced break after each of the first two.

After a total of 21 minutes of listening, subjects received the two-alternative forced-choice test. On each trial, a target and a foil were presented (in counter-balanced order) with a 500 ms pause in between. For each item, subjects were asked to indicate which of the two strings sounded more like a word from the language by pressing either the '1' or '2' key on the keyboard. For each subject, the foils were either nonwords (N = 50) or part-words (N = 50). The order of the test trials was randomized individually for each subject. After all critical trials were presented, the six catch trials were presented in a random order. There was no overt break between the critical trials and catch trials. Placing the catch trials at the end ensured that learning during the catch trials could not affect the critical trials, preserving the comparison with the original study.

Participants also completed an additional demographic survey regarding age, hearing, and native language.

Exp. 1b: University+MBROLA

Subjects in Exp. 1b were undergraduates recruited and tested individually at Boston College in a laboratory testing room. Stimuli were presented with a MacBook laptop and Dell AX510 speakers. 62 subjects participated in the part-word condition (Ages 18–22, M = 19), and 64 subjects in the nonword condition (Ages 17–21, M = 19). All but 10 subjects were native speakers of English. Recruitment continued until we had 50 subjects who passed the catch trials in the part-word condition (Ages 18–21, M = 19) and 50 in the non-word condition (Ages 17–21, M = 19). Thus, Exp. 1b had a somewhat larger sample than the other experiments in this paper.[4] Otherwise, Exp. 1b was identical to Exp. 1a.

Exp. 1b was formally preregistered. Time constraints did not permit formal preregistration of the other experiments. Because faithful replications are tightly constrained by the published originals, we did not prioritize formal preregistration.

Exp. 1c: AMT+WatsonTTS

Subjects. 101 native English speakers were recruited and tested through Amazon Mechanical Turk: 50 subjects in the part-word condition (Ages 21–56, M = 33) and 51 subjects in the nonword condition (Ages 20–59, M = 36). The additional subject in the nonword condition was due to software error. There were no catch trials and thus no exclusions.[5]

Materials. Training materials were generated analogously to those in Exp. 1a & 1b, but using IBM Watson's Text to Speech with the American English voice *Allison* (IBM, 2017). In order to generate natural-sounding speech, WatsonTTS uses variable prosody and intonation. In order to create relatively monotone speech, we produced each syllable in isolation (and thus without co-articulation between syllables). We edited each sound file to 0.277 sec in length. Because the Watson text-to-speech system does not allow one to specify specific lengths for phonemes or syllables, some of the syllable sound files had small amounts of silence at the beginning and/or end.

The resulting sound files were then concatenated using Sound eXchange (SoX) to make 252 tokens of each of the 6 words in a random order, with the stipulation that the same word never occurred twice in a row (Bagwell & Contributors, 2015). The resulting sound file was converted from WAV to MP3 using SoX. Expected transitional probability within a word ranged from 42.5% to 100% and ranged between words from 5% to 60% (unfortunately, the exact numbers were lost and difficult to recover from the raw audio).

Test trials were constructed as in Exps. 1a & 1b, though using different foils. (Exp. 1c was historically the first to be run, and given its poor results, we generated new items before running Exps. 1a & 1b.) The nonwords were *babita, dabibi, dudata, pudata, tadupu, tutapi* and the part-words were *babubi, pidata, patubu, bitaba, budabu, tadabu*. Of the six part-words, the first three matched a trained word on the first two syllables and the last three matched a trained word on the final two syllables. As in Exps. 1a & 1b, each test trial consisted of a trained word and a foil, with an approximately 500 ms gap in between (the exact length varied slightly because the 227 ms sound files for each syllable contained a variable amount of silence at the beginning and end. See above.). There were no catch trials.

Procedure. The procedure matched that of Exp. 1a, with the exception that due to a programming error, a single order of test trials was used for all subjects.

Comparison with original. The reproduction of SNA96-1 is complicated by some inconsistencies in the

original paper. For instance, in describing the training, it provides mutually incompatible numbers: 300 tokens of each of the 6 words at 216 syllables/minute for a total of 21 minutes and 4,536 syllables. Similarly, they state that the maximum between-word transition probability was 20%, whereas their description of the stimuli indicates that this number must have been greater than 50%. Another notable confusion is that SNA96-1 states that the language consisted of 12 syllables, but the list of words makes use of only 11. There were other, smaller inconsistencies, detailed in the preprints. The method described above represents what we believe to be the most sensible reconciliation of these inconsistencies (for calculations and discussion, see Garcia, Iozzo, et al., 2017).

Otherwise, the most notable differences between the replications and the original are the ones specifically manipulated across the three replications: the subject pool (AMT vs. university subject pool), screening of innatentive subjects (which the original did not do), and the using MBROLA or WatsonTTS instead of MacinTalk.

There are several other more minor differences. For instance, the original familiarized subjects with how to use a computer keyboard, which we judged to be unnecessary, particularly for users of Amazon Mechanical Turk. Similarly, the original enforced a five minute break between training segments, whereas ours was self-paced. An exhaustive comparison of the methods in the three replications and the original can be found in the preprints (Garcia, Iozzo, et al., 2017; Garcia, van Horne, & Hartshorne, 2017; Hartshorne, 2017; Hartshorne & Skorb, 2018; Iozzo et al., 2017; Mu et al., 2017; Murdock et al., 2017a, b).

Results

The description of the results below omits some details for reasons of space and readability. See the preprints for an exhaustive description of the results (e.g., every test statistic, standard error, and model specification).

Catch trials. The comparison of catch trial accuracy for our Amazon Mechanical Turk subjects (Exp. 1a) and Boston College students (Exp. 1b) is shown in **Figure 1**. The latter did significantly better than the former (0.94 vs. 0.83; Wald's $z = 5.96$, $p = 2.5*10^{-9}$).[6] As shown below, however, excluding subjects who missed catch trials had a negligible effect on the pattern of results.

Learning Success. The original reported significant, above-chance discrimination of trained words from foils in both the nonword condition (76%) and partword condition (65%). This was confirmed in Exps. 1a and 1b, whether or not inattentive subjects were excluded (**Table 1**). The results for Exp. 1c were more mixed: results were in the right direction but only significant under the original (anti-conservative) analyses, but not the revised analyses.

As with the catch trials, a direct comparison of the AMT sample and the lab-based sample using the same materials (Exp. 1a vs. Exp. 1b) revealed slighty but significantly better performance in the latter, even after removing inattentive subjects (Wald's $z = 3.08$, $p = .002$; **Figure 2**).

Nonword vs. part-word foils. Whereas the original reported significantly better accuracy in the nonword condition vs. the partword condition, this was not observed in any of the three replications, whether or not inattentive subjects were removed (**Table 1**).

High- and Low-Transition probability stimuli. The six trained words varied in their internal transition probabilities (see Method). The original reported significantly better discrimination of the three trained words with the highest internal transition probabilities vs. the three trainned words with the lowest internal transition probabilities. This was not observed in any of our replications, whether or not inattentive subjects were excluded (**Table 1**).

Partwords distinguished by first syllable vs. final syllable. Of the six partwords, three differed

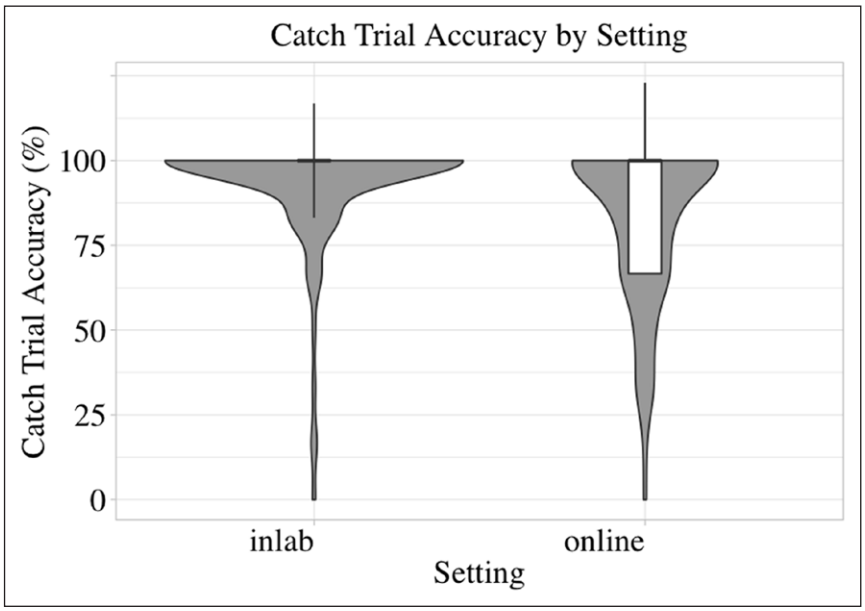

Figure 1: Violin plot of catch trial performance for all subjects in Exp. 1a (AMT+MBROLA) and Exp. 1b (Uni+MBROLA). Note that the boxes represent the interquartile range, and the whiskers represent standard deviation.

Table 1: Effect Sizes Across Replications and Original (SNA96-1).

Study	Synth	Venue	Exclude Inattentives	Nonword	Partword	Bupabi	Non- v. Partword	High- vs. LowTP	Change-first v. -last
Original	MacinTalk	Uni	no	76**	65**	74***	11*	7*	19**
Exp. 1a	MBROLA	AMT	no	61***	57***	52	4	−1	1
Exp. 1a	MBROLA	AMT	yes	65***	62*	57	3	0	5
Exp. 1b	MBROLA	Uni	no	67***	65***	68***	2	2	1
Exp. 1b	MBROLA	Uni	yes	70***	68***	71***	2	2	2
Exp. 1c	WatsonTTS	AMT	no	57	58	NA	−1	−1	−8

Note: Accuracies for nonword condition, partword condition, and the item *bupabi,* as well as differences in accuracies for the three main comparisons reported in SNA96-1. Green represents a successfully replicated effect across all replications. Standard errors, test statistics, and exact p-values are reported in the preprints.

* Significant at alpha level p = .05 under the revised analyses, [**] p = .01 under revised analyses, [***] p = .001 under revised analyses.

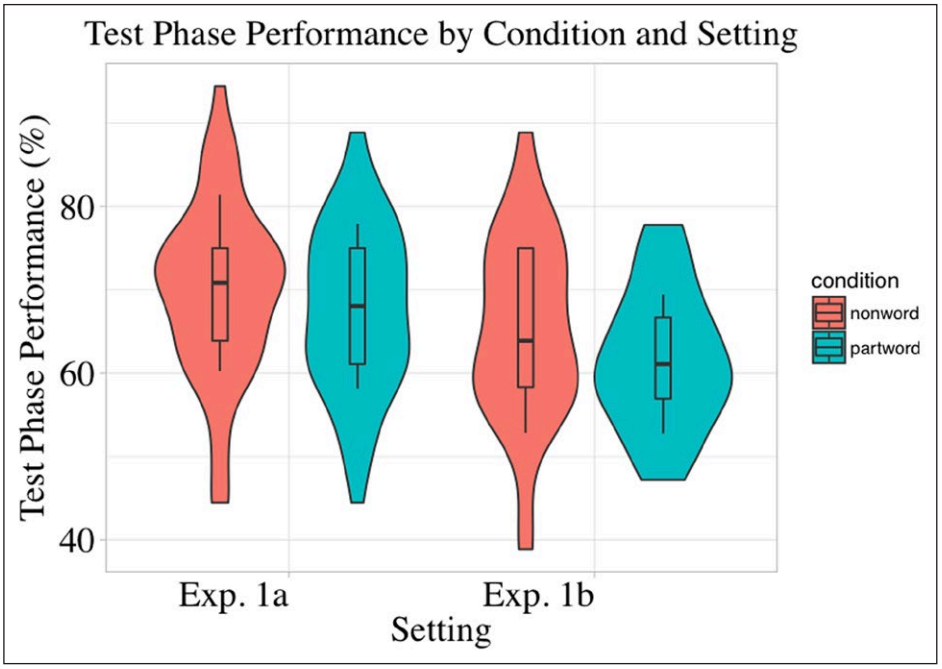

Figure 2: Comparison of attentive subjects only in Exp. 1a (AMT+MBROLA) and Exp. 1b (LAB+MBROLA), by condition. Note that the boxes represent the interquartile range, and the whiskers represent standard deviation.

from trained words on their first syllable, and three on their final syllable (see Method). The original reported a significantly lower false alarm rate for the former relative to the latter. This was not observed in any of our replications, whether or not inattentive subjects were excluded (**Table 1**).

Bupabi. The most common syllable pair in the training was *bupa*. The original reported that accuracy in rejecting the partword foil *bupabi* was particularly high (74%) and significantly different from chance, which the authors interpreted as further evidence for a direct effect of transition probabilities. Our results were mixed, with performance differing from chance in Exp. 1b (Uni+MBROLA) but not Exp. 1a (AMT+MBROLA) (**Table 1**). Exp. 1c did not include a partword beginning with *bupa*, and so this analysis could not be conducted for that experiment.

Discussion of Exp. 1

Across three replications, we saw consistent evidence of adults' ability to successfully engage in statistical word segmentation, confirming a major finding of SNA96-1. However, we saw no evidence of the three major moderators they investigated: better rejection of nonword vs. partword foils, better recognition of high- vs. low-transition probability words, and better rejection of foils that differ from trained words on their first syllable than last syllable. We saw only mixed evidence regarding their finding of particularly successful rejection of the foil *bupabi*. This was a decidedly mixed outcome.

Our three replications allowed us to test the importance of some of our methodological decisions. As expected, the choice of synthesizer (MBROLA vs. the more modern WatsonTTS), subject population (AMT vs. university undergraduate subject pool), and attentional screen had

minimal effect on the absolute results and no effect on the pattern of significance. The one exception to this was analysis of the foil *bupabi*, where the direction of results was the same across replications but the significance varied.

Overall learning and catch trial accuracy were better in the university sample than the AMT sample. Given that this remained even after a fairly difficult attention screen, it likely reflects the greater education, higher IQ, and higher SES of the university sample – all factors that tend to correlate with doing better on tests. In any case, the numerical effect was small and did not translate into a difference in the pattern of results (with the caveat about *bupabi*). Based on these findings, and the enormous literature reporting good results from AMT subjects, we use AMT for the remaining replications.

Replication of Saffran, Johnson, Aslin, & Newport (1999) Exp. 1

The original experiment extended the results of nonword condition from Exp. 1 of Saffran, Newport, & Aslin (1996) to languages consisting of musical notes rather than speech syllables. As described below, the design of the study matched the 1996 study very closely. One notable improvement is that subjects were exposed to one of two languages constructed such that the trained words for half the subjects served as the foils for the other half. All foils were "nonwords": a sequence of three tones that had 0 transitional probability in the trained language. The pattern of results was largely the same as in the 1996 paper, indicating that statistical word segmentation is not limited to speech sounds. Again, the authors found better recognition for trained words with higher internal transition probability, suggesting a crucial role for transitional probability in learning. The converging results between this experiment and the original lent greater credence to both.

Method

The preprint describing the replication in detail can be found at psyarxiv.com/qmptz.

Subjects. 99 subjects were recruited through Amazon Mechanical Turk and paid for their participation. Subjects were randomly assigned to one of the two languages, with 50 subjects in Language One and 49 in Language Two. (We intended to recruit 50 for each, but ended up one short due to experimenter error.) Following the original, we excluded five subjects who reported formal music activities (lessons, singing in a choir, etc.) in the last five years. We further excluded 17 who incorrectly answered any catch trials, leaving us with 77 subjects. Note that based on the results of Exp. 1, we did not perform analyses including the inattentive subjects in any of the subsequent replications. However, we included the inattentive subjects with the published data, so the reader may perform these (and other) analyses if desired. The remaining sample contained 39 subjects in Language One (mean age: 34, range: 21–56) and 38 in Language Two (mean age: 35, range: 24–65). All but one of these subjects were native speakers of English, and all reported normal hearing.

Materials. Tones were constructed out of eleven pure tones of the same octave (starting at middle C within a chromatic set) and were the same length (.33 s). In order to form six tone words, the pure tones were combined into groups of three. Language One consists of: ADB, DFE, GG#A, FCF#, D#ED, and CC#D. Some tones appeared in only one word (G# for example), while others occurred in multiple words (D for example). The six tone words were then used to create six different blocks that each contained 18 words. The tone words were concatenated in a random order without silence between words, and tone words did not occur twice in a row. Each word was concatenated together 70 times in order to produce a seven-minute continuous stream of tones, which was called Language One. A given tone stream, for example, would be DFEFCF#CC#DD#EDGG#A. There were no acoustic markers of word boundaries. Language Two was constructed in the same manner as the first, using the same original eleven tones but combined in different order to form six new words. Language Two consists of: AC#E, F#G#E, GCD#, C#BA, C#FD, G#BA.

Within both languages, transitional probabilities alone were the only consistent cues to the beginnings and ends of the tone words. In terms of Language One, the average transitional probability between tones within words was 0.64 (range: 0.42–1.00). In contrast, the average transitional probability between tones spanning word boundaries was 0.18 (range: 0.04–0.59). The transitional probabilities between the tones in Language Two averaged .71 (range 0.41–1.00), with lower average probabilities across word boundaries (M = 0.21, range: 0.04–0.49).

A thirty-six item test was constructed in which participants were forced to select between two alternative choices in order to assess learning. Each test item was made up of two tone sequences: a word and a nonword. Nonwords consisted of three-tone sequences that were also made using tones drawn from the language, but never occurred during exposure to the language (and thus had transitional probabilities of 0). One of the sequences presented on each trial was a word from Language One and the other was a word from Language Two. Thus, the trained words for subjects learning one language were the foils for subjects learning the other. All six words from each of the two languages were paired for a total of 36 test trials. The tone sequences of each trial were separated by a 0.75 second interval of silence.

We also added six catch trials that compared a trained word against something that was obviously not in the training. They were always the last six items in the test.

Procedure. Subjects listened to the seven-minute-long recording of one of the two tone streams (Language One or Language Two) described above, repeated three times. Each of the three seven-minute listening sessions was followed by a short, subject-paced break. After a total of twenty-one minutes of listening, subjects were exposed to the 36-question test. They were instructed to indicate the most familiar tone sequence on each test trial. The correct choice for subjects exposed to Language One was the incorrect choice for subjects exposed to Language Two. The order of the test trials was randomized for each subject.

After the primary test trials, the catch trials followed immediately, without any overt distinction. Having them appear after the critical trials avoided any possibility of contamination of the critical trials. The order of the catch trials was randomized for each subject.

Participants also completed an additional demographic survey regarding age, hearing, native language, and music experience.

Comparison to Original. The largest difference between the replication and the original is the subject population: Although the original does not report their source of subjects, it was presumably not AMT. Because training was randomly generated according to rule, the transition probabilities were very slightly different, with a somewhat clearer delineation of words in our stimuli (see Murdock et al., 2017a). There are a number of other minor differences detailed in the preprint, such as the fact that our subjects responded by keyboard rather than with paper-and-pencil (see Murdock et al., 2017a).

Results

The original reported no significant differences between Language One and Language Two, but did report significantly better performance on high transition-probability words relative to low transition-probability words. In our data, mean scores were likewise similar for Language One (M = 67%, SE = 2%) and Language Two (M = 68%, SE = 2%). The difference between high transition-probability words and low transition-probability words, however, was small for both Language One (High: M = 69%, SE = 3.2%; Low: M = 65%, SE = 2.6%) and Language Two (High: M = 68%, SE = 2.7%; Low: M = 68%, SE = 2.5%).

We assessed statistical significance with a binomial mixed effects regression with main effects of language and transitional probability (high/low) as well as their interaction. We included a random effects for each subject and for each word/foil pair, as well as a random slope of transitional probability (high/low) for each subject (for discussion and analysis of this random effects structure, see Murdock et al., 2017a). We found a significant intercept (B = 0.82, SE = 0.11; Wald's z = 7.8, $p = 8.5^{*}10^{-15}$), indicating overall above-chance learning. The main effect of language was not significant (B = 0.02, SE = 0.09; Wald's z = 0.20, p = .84), nor was the main effect of transitional probability (B = 0.07, SE = 0.06; Wald's z = 1.15, p = .25). The interaction was likewise not significant (B = 0.09, SE = 0.06; Wald's z = 1.51, p = .13). Thus, we replicated the finding of equivalent, significant learning in both languages, but did not replicate the finding of an effect of word-internal transitional probability.

We followed these analyses with binomial mixed effects models for each trained word individually, with a random intercept of foil. Matching the findings of the original, the intercepts were significant for every word but one (ps < .05). However, their exception (ABD) was different from ours (FCF#), suggesting that this lack of significance was not due to the trained word itself.

The original notes that the accuracy they observed for tones (Lang 1: 74%, Lang 2: 80%) was not significantly different from what was reported in SNA96-1 (76%). The accuracy we observe is considerably lower (Lang 1: 67%, Lang 2: 68%), despite our stronger attentional filter. Nonetheless, it is very similar to what we observed in the most analagous replication: 65% (Exp. 1a, with inattentive subjects excluded). Due to the relatively close matches, we did not perform statistical analyses. Because the "words" in Lang 1 were modeled after those in the language used in SNA96-1, the original performed a by-item correlation. Because of the extremely low power (6 items), we did not conduct this analysis.

Discussion

While we replicated Saffran et al.'s (1999) finding of successful statistical word segmentation of tone sequences in adults, learning was strikingly weaker. We also failed to replicate their finding of superior learning for trained words with high transitional probabilities, consistent with our similar failure in Exp. 1.

Replication of Saffran, Johnson, Aslin, & Newport (1999) Exp. 2

This experiment closely mirrored experiment 1 from the same paper, discussed immediately above. However, the words in the two languages were constructed so that they served as "part-word" foils for one another. As in SNA96-1, foils differed from a trained word either in their first tone or in their final tone.

The authors again reported findings in line with their 1996 results: successful discrimination of targets from foils, which was greater when the trained word had higher than average transitional probability or when the foil differed from a trained word on the first tone rather than the final tone. The converging results between this and the two original studies lent greater credence to each.

Method

The preprint for this replication can be found at psyarxiv.com/pj7fb.

Subjects. 101 subjects were recruited through Amazon Mechanical Turk and paid for their participation. Subjects were randomly assigned to one of the two languages, with 51 subjects in Language One and 50 in Language Two. (We intended to recruit 50 for each, but ended up one over due to experimenter error.) Following the original, we excluded twelve subjects who reported formal music activities (lessons, singing in a choir, etc.) in the last five years. We further excluded 17 who incorrectly answered any catch trials, leaving us with 72 subjects: 36 in Language One (mean age: 35, range: 23–59) and 36 in Language Two (mean age: 34, range: 21–59). All but one of these subjects were native speakers of English. All reported normal hearing.

Materials. Language One was identical to that in "Replication of Saffran, Johnson, Aslin, & Newport (1999) Exp. 1". The other language (which we dub "Language Three" to prevent confusion with the "Language Two" of the replication in "Replication of Saffran, Johnson, Aslin, & Newport (1999) Exp. 1") likewise consisted of three-tone sequences made of the same 11 tones. However, it was constructed such that its words were part-words with

respect to Language One, while words from Language One were part-words with respect to Language Three. Note that to generate a part-word, either the first or third tone was substituted with a different tone (e.g., G#DB) such that the resulting transition probability is 0 with respect to the trained language. Three part-words contained the first two tones of words plus a new third tone, and three contained the final two tones of words plus a new first tone. The tone words for the new Language Three were G#DB, DFF#, FG#A, C#CF#, D#EG#, and CC#B. The average within-word transition probability was 0.56 (0.42–1.00). The average transitional probability between words in Language Three was 0.18 (0.05–0.40).

During the test phase, subjects were required to discriminate trained words from foils. The six trained words from one language were paired exhaustively with six words from the untrained language, forming 36 test trials. Because these foils shared some structure with the trained words, this was expected to be relatively difficult. We also added six catch trials that compared a trained word against a foil involving at least one tone from outside the set of 11. They were always the last six items in the test.

The tone sequences of each trial were separated by a 0.75 second interval of silence, and the trials themselves were separated by 1 second.

Procedure. Subjects listened to the seven-minute-long recording of one of the two tone streams (Language One or Language Three) described above, repeated three times. Each of the three seven-minute listening sessions was followed by a short, subject-paced break. After a total of twenty-one minutes of listening, subjects were exposed to the 36-question test. They were instructed to indicate the most familiar tone sequence on each test trial. The correct choice for subjects exposed to Language One was the incorrect choice for subjects exposed to Language Three. The order was randomized for each subject.

After the primary test trials, all subjects answered the 6 catch trials. The order was randomized for each subject. The catch trials followed the primary test trials immediately, without any overt distinction.

Participants also completed an additional demographic survey regarding age, hearing, native language, and music experience.

Comparison with Original. The differences between the replication and original are minor and analogous to those for Exp. 2, and are detailed in the preprint (Murdock et al., 2017b).

Results

Mean scores were 61% (SE = 3%) for Language One and 62% (SE = 2%) for Language Three. Accuracy was slightly higher for foils where the first tone had been changed (63%, SE = 2%) compared with where the last tone had been changed (61%, SE = 2%).

We analyzed the data using a binomial mixed effects model with main effects of language and foil type (changed-first/changed-last), as well as their interaction. We set the contrasts to center the variables, such that the intercept is equal to the grand mean. Again, random effects structure included intercepts for each word/foil pair and for each subject, as well as a random slope of foil type for each subject (for discussion and analysis, see preprint).

We found a significant intercept (0.51, SE = 0.08; Wald's z = 6.0, p = $1.6*10^{-9}$), indicating overall above-chance learning. The main effect of language was not significant (B = 0.02, SE = .07; Wald's z = 0.20, p = .84), nor was the main effect of foil type (B = –0.07, SE = 0.06; Wald's z = 1.3, p = .19). The interaction was likewise not significant (B = 0.09, SE = 0.6; Wald's z = 1.5, p = .15). Thus, we replicated the finding of equivalent, above-chance learning in the two languages, but not the finding of better discrimination of foils where the first tone was changed.

We followed these analyses with binomial mixed effects models for each trained word individually, with a random intercept of foil. Exactly half of the trained words exhibited above-chance discrimination (ps < .05). Despite greater statistical power, this was slightly worse than in the original, where 2/3 were above chance.

Saffran et al. (1999) found that accuracy was lower in their Exp. 2 than in their Exp. 1. We found a similar difference between our replications of their two experiments: M = 67% (SE = 2%) vs. M = 62% (SE = 2%) (B = 0.28, SE = 0.13; Wald's z = 2.2, p = .03).

As in the previous experiment, the original reports a number of follow-up analyses that we did not attempt to replicate. A number involve comparisons to an analogous experiment in (Saffran, Newport, & Aslin, 1996). As we have not yet replicated this study, we could not do these comparisons. The others involved analyses of the effects of the tonal structure of the stimuli, none of which had revealed significant results.

Discussion

Again, we replicated evidence of statistical segmentation of tone sequences, but we did not replicate the theoretically-important moderator. This matches our pattern of replication for the analogous experiment with speech stimuli (see "Investigation of the method: Three replications of Saffran, Newport, & Aslin (1996) Exp. 1").

Replication of Finn & Hudson Kam (2008) Exp. 1

This experiment asked to what degree learners are guided by the phonotactics of their native language when engaging in statistical word segmentation. Each word in the novel language began with a consonant cluster. For subjects in the experimental condition, this consonant cluster was illicit in English. The authors hypothesized that this would result in mis-parsing the input. This was assessed by seeing whether the subjects could discriminate trained words from foils that lacked the illicit initial consonant cluster. Crucially, the foils consisted of phoneme sequences that occurred in the input, resulting from parsing the input with respect to English phonotactics. The authors found that subjects failed in this discrimination, though they did show evidence of learning in other respects (subjects did not simply tune out). The results of a control condition, which used licit onset consonant clusters, ruled out the possibility that learners simply could not handle

consonant clusters. The results of this study suggested a possible avenue for understanding difficulties in second language learning.

Method

The preprint describing this replication can be found at psyarxiv.com/2xcwk.

Subjects. 100 subjects were recruited through Amazon Mechanical Turk and paid for their participation. Whereas in the prior experiments, we required subjects to get all six catch trials, for this experiment we had eight (reflecting the number of trained words in the different experiments). We elected to allow up to one mistake. 29 subjects were excluded for missing more than one catch trial, leaving 37 in the control condition (Ages 23–61, M = 33) and 34 in the experimental condition (Ages 24–63, M = 37). All but one of the remaining subjects were native English speakers, and all reported normal hearing.

Materials. We were unable to recreate the original materials, which were described using a phonetic alphabet specific to a now-obsolete speech synthesizer, for which we were unable to locate documentation. Given that the phonetic structure of the stimuli was key to the experiment, we were not comfortable guessing. The authors graciously provided us with the original stimuli.

Training. Experimental and control stimuli both consisted of eight two-syllable words (CCVCV), each beginning with a consonant cluster. For the experimental stimuli, these CC onsets violate the word-initial phonotactic rules of English. In the control stimuli, CC onsets are licit. The authors report that they generated the stimuli with the text-to-speech program SoftVoice, producing syllables with a monotonic F0 (fundamental frequency) of 83.62 Hz, no co-articulation effects, and matching vowels for length.

During training, words were presented quasi-randomly with no pauses and no immediate repetitions. The authors report transitional probabilities of 1.0 for syllable transitions that are word-internal and .143 at word boundaries. They further report phoneme transitional probabilities (PTPs) within words were higher than those across word boundaries, with word-internal PTPs, ranging from .25 to 1.0 and PTPs across the word boundaries ranging from .035 to .143.

Training lasted 17 min 59 sec. While the original report that each word occurred 560 times, we suspect this is a typo, since it would suggest 8.3 syllables/sec. The speech rate in the training files provided is much closer to 1 syllable/sec, which would be consistent with each word occurring 70 times for a *total* of 560 words in the training. However, we did not confirm this with an exact count of the words.

Test. After exposure, participants were given a forced-choice test between a trained word and either a nonword (consisting of two syllables from the language but a transitional probability of 0) or a split-word (a trained word minus the first consonant and with another trained word's initial consonant at the end, resulting in a CVCVC structure). In essence, the researchers constructed the split-cluster words by shifting one phoneme to the right in the exposure stimuli. Note that in the experimental condition, this results in a foil that is a licit English word, pitted against a target which is not. The split-word foils measured the accuracy of parsing the consonant clusters in defiance of English phonotactics, while the nonword foils served as a (high) baseline.

There were 8 of each type of test item (word vs. nonword, word vs. split-cluster word), yielding 16 test items in total for each language. The two items in a pair were presented one after another with a 1 second pause in between. There was a 500 ms pause in between pairs during which participants were expected to answer. The order of test trials was randomized for each subject.

Upon inspection, two of the nonwords (one in the control condition and one in the experimental condition) did not have an audible vowel in the first syllable, resulting in a CCCV structure. In principle, this should have made it easy for the subjects to reject the nonword, though in fact accuracy was high for one (0.89) but not the other (0.53). In order to facilitate comparison with the original, the numbers reported below include those items. Cases where exclusion affects the pattern of results are noted below. For full details, see the preprint.

In addition to these trials, we included eight catch trials that pitted a trained word against a nonword that included phonemes that never appeared in the training. These nonwords were created using MBROLA (Dutoit et al., 1996). Due to experimenter error, one of the catch trials in the experimental condition did not have a correct answer, and so was excluded. The catch trials were always the last eight items in the test, and the order of the catch trials was randomized for each subject.

Procedure. Subjects listened to the training stimuli in two separate listening blocks, with a self-paced break in between. They were encouraged to scribble or color with markers or pens during exposure. In order to determine whether they followed this instructions, at the end of the experiment they were asked to describe their drawing. Six of the subjects reported forgetting to draw anything. Excluding these subjects had minimal effect on the results (see experiment preprint), so we include them in the analyses reported below. Because the original does not report whether subjects complied with the drawing instructions, we do not know whether this rate differs from that in the original.

After exposure, participants were given the forced-choice test described above. They were instructed to listen to pairs of possible words and were asked to choose which word was a better example of the language. Participants indicated responses using two keys on their keyboard, pressing '1' if the first item in the pair sounded closer to the language and '2' if the second one sounded closer to the language.

Participants also completed an additional demographic survey regarding age, hearing, and native language.

Comparison to Original. The most significant differences were the subject population and the use of catch trials. Other minor differences are reported in the preprint, such as the fact that we randomized the order of test trials rather than using two fixed orders and

used a shorter inter-trial interval (Garcia, van Horne, & Hartshorne, 2017). Note that in contrast to most of the other replications in this paper, the stimuli were not merely *as described* in the original, but were the actual original stimuli.

Results

The original reported above-chance learning in all cases *except* the split-word items in the experimental condition. Moreover, they found that performance on the split-word items was significantly worse in the experimental condition relative to the control condition.

Mean performance on the nonword items was 71% (SE = 3%) in the control condition and 62% (SE = 3%) in the experimental condition. Performance on the split-word test was 62% (SE = 3%) in the control condition and 56% (SE = 3%) in the experimental condition. We modeled these data with fixed effects for condition and foil type (nonword vs. split-word) and their interaction. We included random intercepts both for subject and for target/foil pair, plus a random slope of test type (nonword vs. split-word) by subject (for details, see the preprint). We set the contrast structure to center the variables, which results in the intercept being the grand mean.

We found overall significant learning, reflected in a significant intercept (B = .56, Wald's z = 5.2, $p = 2.3*10^{-7}$). However, neither of the main effects nor the interaction were significant (ps > .1). Follow-up analyses looking at the each condition and foil-type individually revealed significant intercepts for nonwords in the control condition (Wald's z = 4.2, p = .00004), split-clusters in the control condition (Wald's z = 2.6, p = .008), nonwords in the experimental condition (Wald's z = 2.4, p = .02), but not split-clusters in the experimental condition (Wald's z = 1.1, p = .28).

The pattern of significance was slightly different when using the same statistical tests as were deployed in the original. Evidence for above-chance learning in the split-word test in the experimental condition as only marginal (p-values varied between .05 and .10, depending on whether the bad items and non-drawers were included or excluded). The original reports no difference between the control and experimental conditions for nonword tests, whereas in our replication of their analyses, this comparison was sometimes significant, depending on whether the two bad items or the non-drawers were included. For full presentation of these analyses, see experiment preprint.

Discussion

The primary theoretically-relevant finding reported in the original was that English-speaking adults failed to successfully segment words that violated the phonotactic constraints of English but could successfully segment phonotactically-licit words. However, subjects could reject foils that had a between-syllable transition probability of 0 equally well in both conditions.

In contrast, we did not observe significant differences between conditions. Thus, the original conclusions are not supported. The one caveat is that when we compared each of the four trial types against chance, we sometimes replicated the pattern observed in the original — depending on which method of analysis was used. Thus, in terms of replicating statistical patterns, the results are mixed. This mixed set of results could be consistent with insufficient power (though note that our replications were higher-powered than the original). In ongoing work we are running a higher-powered replication.

Replication of Finn & Hudson Kam (2008) Exp. 3

This experiment repeated the experimental condition of Exp. 1 from the same paper (described above) with a slight change in the instructions: Subjects were explicitly presented with one of the words from the language prior to training ('kmodu'). The hypothesis was that this might help learners succeed in learning the words with illicit onset consonent clusters. It did not. The converging results lent greater credence to both the follow-up and the original.

Method

The preprint describing this replication can be found at psyarxiv.com/ekw9c.

Subjects. 50 subjects were recruited through Amazon Mechanical Turk and paid for their participation. Following our replication of Exp. 1, we excluded 11 subjects who incorrectly answered more than one catch trial, leaving 39 (Ages 23–61, M = 33). All subjects were native speakers of English and reported normal hearing.

Materials and Procedure. Materials and procedure were identical to those in the Experimental condition in "Replication of Finn & Hudson Kam (2008) Exp. 1" with the following exception: After the more general instructions participants were told, "*Kmodu* is a word in the language you are about to listen to." This information was presented in written form, and the sound file for *kmodu* was also played. Before beginning training, the subjects were again informed that *kmodu* is a word in the language.

Comparison with Original. In addition to the population (AMT vs. University) and use of catch trials, the other difference worth mentioning was that we familiarized the subjects with *kmodu* via recording, rather than with a live experimenter. The former has the advantages of avoiding any experimenter bias but may be less salient to the subjects. Otherwise, the differences were extremely minor and analogous to those in Replication of Finn & Hudson Kam (2008) Exp. 1 (Iozzo et al., 2017).

Results

Mean accuracy was 62% (SE = 8%) for the nonword foils and 58% (SE = 8%) for the split-word foils. The original reports that performance was above-chance for the former and at chance for the latter (they do not report a comparison of the two).

We analyzed the data with a binomial mixed effects model with a fixed effect for foil type (nonword vs. split-word) and random intercepts both for subject and for target/foil pair, plus a random slope of test type

(nonword vs. split-word) by subject. We set the contrast structure to center the fixed effects, which results in the intercept being the grand mean. We found a significant intercept, reflecting overall successful learning of the trained words (B = .64, Wald's z = 2.65, p = .008). The main effect of foil type was not significant (B = –0.09, Wald's z = 0.38, p = .70). We also ran models with each foil type separately, finding significantly above-chance learning for nonword trials (B = .73, Wald's z = 2.30, p = .02) but not split-word trials (B = .55, Wald's z = 1.54, p = .12). Interestingly, when we replicated the original analyses, we find significant results in both cases. Removal of the bad item, the non-drawers, or both did not qualitatively affect any of these results (see experiment preprint). Note that as in the previous experiment, we do not know how the percentage of non-drawers in our experiment (6%) compares with the original, for which the numbers were not reported.

The original reports ceiling accuracy (100%) on trials involving *kmodu* itself. We find a somewhat lower rate of 83% (SE = 4%).

Discussion

The results of our replication of Exp. 2 mirror those of our replication of Exp. 1: We replicate the difference in significance between conditions using the revised analyses but not the original analyses. However, a difference in significance is not a significant difference, and neither set of analyses finds a significant difference. Strictly speaking, this last finding is not a failure to replicate, because the original did not test for a significant difference. However, since their interpretation depends crucially on there being a significant difference between conditions, it seems reasonable to consider our failure to find such a difference to be a failure to replicate.

As was the case in our Exp. 1, the numerical pattern was in the expected direction. Thus we cannot rule out the existence of a relatively small effect in the expected direction. We are currently testing this with a higher-powered replication.

Replication of Frank, Goldwater, Griffiths, & Tenenbaum (2010) Exp. 1

The languages in the above experiments lacked a number of common linguistic features, such as sentence structure or words of varying length. This experiment investigated the effects of varying word and sentence length on statistical word segmentation, finding that longer words were no harder to learn but longer sentence were harder to learn from. These results, along with results from other experiments in this paper, were then used to compare five prominent models of word segmentation, finding that two of them failed to meaningfully account for the effects of sentence length. The authors do not report evaluating whether the models accounted for the (lack of) word length effects.

Method

The preprint describing this replication can be found at psyarxiv.com/79bnu.

Subjects. We recruited 400 subjects from Amazon Mechanical Turk, randomly allocating 50 subjects to each of eight sentence length conditions. Within each sentence length condition, each subject was randomly assigned a unique language out of 50 total languages. In other words, 50 languages were reused across sentence length conditions, but no language was used twice within a sentence length condition.

For analysis, we dropped 6 subjects whose native language was not English and 109 participants who missed more than one out of six catch trials, leaving 285 participants, which were still reasonably evenly distributed across languages and sentence length conditions (32 – 39 subjects/condition). The ages of the subjects ranged from 20 to 73 (M = 35.7).

Materials. Each participant heard a unique and randomly generated sample from one of 50 randomly generated artificial languages. The lexicon of this language was generated by concatenating 18 syllables (/ba/, /bi/, /da/, /du/, /ti/, /tu/, /ka/, /ki/, /la/, /lu/, /gi/, /gu/, /pa/, /pi/, /va/, /vu/, /zi/, /zu/) into six words: two with two syllables, two with three syllables, and two with four syllables. All speech was generated using the MBROLA speech synthesizer (Dutoit et al., 1996) with the us2 diphone database. All consonants and vowels were 25 and 225 ms in duration, respectively, and the fundamental frequency of the speech was 100 Hz. Although the authors helpfully posted their stimuli, we did not use these. We needed many more languages (given our much larger number of subjects), and it was simpler to create all of them rather than merely most.

Training. Subjects were randomly assigned to one of eight sentence length conditions (1, 2, 3, 4, 6, 8, 12, or 24 words per sentence). As training data, subjects heard 600 words, 100 each of the 6 words from the lexicon, broken into sentences of the given sentence length. Sentences were generated by randomly concatenating words with no adjacent word repetition. No breaks were present between words in the sentences (equal co-articulation between every syllable), but there was a 500 ms break between sentences during training. Length of training data ranged from 7.7 (length-24) to 12.5 minutes (length-1); shorter sentence length conditions had longer training materials due to more sentences and thus more breaks.

Testing. Test materials consisted of 36 target-distractor pairs from the language. 30 of the pairs consisted of a word from the language paired with a "part-word" distractor. The part-word contained the same number of syllables as the trained word it was paired with, and was composed of the end of one word and the beginning of another from the language. For each language, 5 part-word foils were randomly generated for each of the 3 word lengths (2, 3, and 4 syllables), resulting in 15 part-words. Except for the length-1 condition, part-word sequences appeared in the sentences of the corpus, albeit at lower frequencies than true words.

Each part-word was then paired exhaustively with trained words of the same length, making 30 target-distractor pairs, each trained word appearing 5 times with same-length part-words, and correspondingly, each

part-word appearing twice with same-length trained words.

To ensure subjects were paying attention to the task, we created 3 catch words, composed of syllables that never appeared in the training data. Like the part-words, each catch word appeared twice with each trained word of the same length, resulting in 6 catch trials and a total of 36 test pairs, which were intermixed during testing.[7]

Test pairs were the same for a specific language regardless of sentence length condition. Test pairs were shuffled so that no trained or part-word occurred twice in a row, and specific pairs shuffled such that word presentation order was random. Each pair was played with a 500 ms break between words during testing.

Procedure. Subjects were instructed to listen to a nonsense language for 15 minutes, and told that they would be tested on how well they learned the words of the language. After listening, they were instructed to make two-option forced-choice decisions between the test pairs indicating which sounded more like a word from the language by pressing either '1' or '2' on the keyboard.

Comparison with Original. Beyond the usual differences (population, use of catch trials), there was a minor difference in how we generated the stimuli: we reused the same 50 languages at each sentence length condition, thereby decreasing scientifically uninteresting random noise. This should not affect the direction of results, but could make it easier for us to detect trends across sentence sizes. This advantage is somewhat mitigated by the fact that we screened out a substantial number of subjects. This change in the stimuli allowed us to treat stimuli as random effects, which the original design did not permit (otherwise, they used well-justified analytic models). A few other minor differences are noted in the preprint (Mu et al., 2017). Note that while we used the same speech synthesizer (MBROLA) as in the original, we actually used a different voice (us2 instead of us3) due to a(n apparently new) bug in the us3 database (Mu et al., 2017).

Participants also completed an additional demographic survey regarding age, hearing, and native language.

Results

Subject performance across sentence length condition is depicted in **Figure 3**. As in the original, we observed significant variation among individual performance in line with classical word segmentation studies (Saffran, Newport, & Aslin, 1996), but a systematic trend in mean participant performance across sentence lengths.

The original analyzed the data with a binomial mixed-effects regression with a continuous fixed effect of word length, a categorical fixed effect of sentence length, and a random effect of participant identity. To this, we added a random effect language, reflecting the fact that we (but not the original) re-used languages across sentence lengths. We compared this to a model with a fixed effect of the interaction of word length and sentence length, and – like the original – we found that this interaction did not improve fit ($\chi^2(7) = 13.3$; $p = 0.06$).[8]

Unlike the original, we observed a negative effect of word length ($B = -0.11$; $p < 0.001$). Intuitively, this means that subjects had a harder time correctly answering questions with longer-length words. Furthermore, unlike the original, which reported significant learning only for the shortest sentence lengths, we observed significant coefficient estimates for all sentence lengths (ps < .01), indicating that performance in all sentence length conditions was significantly higher than chance.

However, these differences in significance patterns are likely to be due to the greater power in our study. We re-analyzed both our data and the original data (which is available online) with a model that treated sentence length as a continuous predictor. In both cases, we again found no evidence for an interaction between word length and

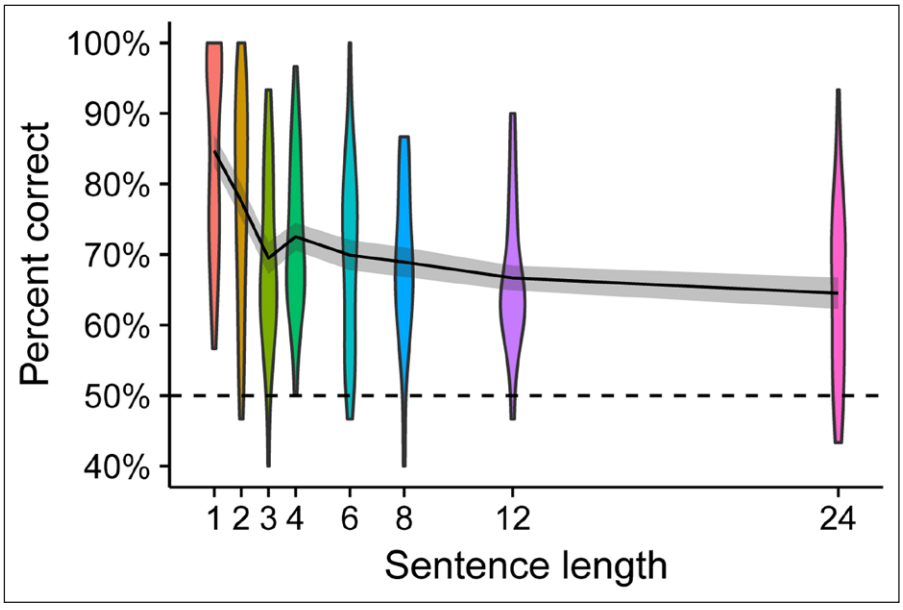

Figure 3: Violin (density) plots of mean participant performance in each sentence length condition. A line connects the means of each group, with shaded area representing standard error. The dotted line represents chance performance.

setnence length and so did not include said interaction term in further analyses (**original:** $\chi^2(1) = 0.0$; $p = 0.96$; **replication:** $\chi^2(1) = 0.0$; $p = 0.98$). In both datasets, we observe a significant negative effect of sentence length (**original:** B = –0.06, SE = 0.012; p = 0.005; Cohen's d = –0.248; **replication:** B = –0.04; SE = 0.006; $p < 0.001$; Cohen's d = –0.147). However, the effect of word length was significant in our dataset but not the original (**original:** B = 0.001, SE = 0.028; p = 0.96; *Cohen's d* = 0.001; **replication:** B = –0.11, SE = 0.012; $p < 0.001$; *Cohen's d* = –0.048).

Discussion

We confirm the original's main conclusion that word segmentation becomes considerably more difficult as sentence length increases. Unlike the original, we find above-chance performance even for the longest sentence lengths. This difference could be due to our greater power or our stricter attention screen. Indeed, the overall shape of **Figure 3** is a remarkable match for the analogous graph in the original. Moreover, regression analyses suggest a similar effect of sentence length in both datasets. Regardless, this difference in pattern of significance does not affect the main conclusions.

The other minor difference in our results is that we observed a significant negative effect of test pair word length on segmentation ability. The original observed only a negligible effect of word length. Given the relatively small size of the effect in our data set, a likely explanation for the differences between studies is the difference in statistical power.

General Discussion

Above, we report replications of six experiments from the adult statistical word segmentation literature. In the remainder of this paper, we address findings with regards to the reproducibility of the adult statistical word segmentation literature, the replicability of the literature, and implications for the field as a whole.

Reproducbility of Adult Statistical Word Segmentation Literature

With regards to reproducibility – the ability to rerun an experiment as described – our results were mixed. While in general the experiments were fairly thoroughly described, three of the experiments (Saffran et al. (1996) Exp. 1; Finn & Hudson Kam (2008) Exps. 1 & 3) had significant errors in the description of their stimuli. Moreover, the description of the stimuli in two of the experiments used an obsolete phonetic description that we were unable to decipher (Finn & Hudson Kam (2008) Exps. 1 & 3). Relatedly, of the six experiments, we were only able to obtain the speech synthesizer used by one of them (Frank et al., 2010). In sum, only one of the six experiments could be straightforwardly reproduced, in part because all of its stimuli, code, and data were publically available (Frank et al., 2010).

None of this is ideal from the perspective of a cumulative science. On the other hand, we were ultimately able to reproduce reasonable facsimiles of all the experiments as described. Thus, we believe these issues represent areas for improvement in our field, rather than fatal flaws.

Note that this conclusion is based on six experiments, and so should be treated as a very coarse estimate. As the project continues and more experiments are replicated, we should have an increasingly clear picture of the literature.

Replicability of the Adult Statistical Word Segmentation Literature

Across all six paradigms — and eight of the nine replications— we find clear evidence of statistical word segmentation in adults. This generalized across a range of stimuli, including non-linguistic "tone words."[9] There does not seem to be much doubt in the literature about whether statistical word segmentation is possible, but if so, this should remove it.

These six experiments reached a number of additional theoretical conclusions based on observed moderations of statistical word segmentation. Arguably the most important are:

1. Words with higher internal transitional probability are easier to recognize.
2. It is easier to reject a foil that matches the end of a trained word than the beginning.
3. Learners privilege the phonotactics of their native language over transitional probabilities when doing statistical word segmentation.
4. The length of the trained word does not affect how easy it is to segment.
5. The longer the sentence, the more difficult statistical word segmentation is.

(1) had provided important evidence for a mechanistic role of transitional probability in word segmentation. Interestingly, (5) suggested that transitional probability was only correlated with the true mechanism (Frank et al., 2010). (3) suggested that second language learners might be impaired in their use of statistical word segmentation due to interference from the phonotactics of their first language. (2) and (4) appear to have played a minimal role in theory.

The differences in replicability for the main effect (successful statistical word segmentation) and the five key moderators above is clearly shown the in the two panels of **Figure 4** (see also **Tables 2** & **3**). While the sizes of the main effects are on average slightly smaller in the replications than the originals, there is a clear correlation. In contrast, there is hardly any relationship between the originally reported moderations and the replications of those moderations.

Indeed, only (5) was fully supported by our replications. We found no evidence for (1) or (2) or (4). Some of the statistical analyses that supported (3) replicated, but not the crucial ones (though there are hints that there may be a very small effect in the expected direction; addressing this question requires and even more highly powered study).

As discussed in the Introduction, accounting for a failure to replicate is notoriously difficult, given that there are inevitably differences between the replication

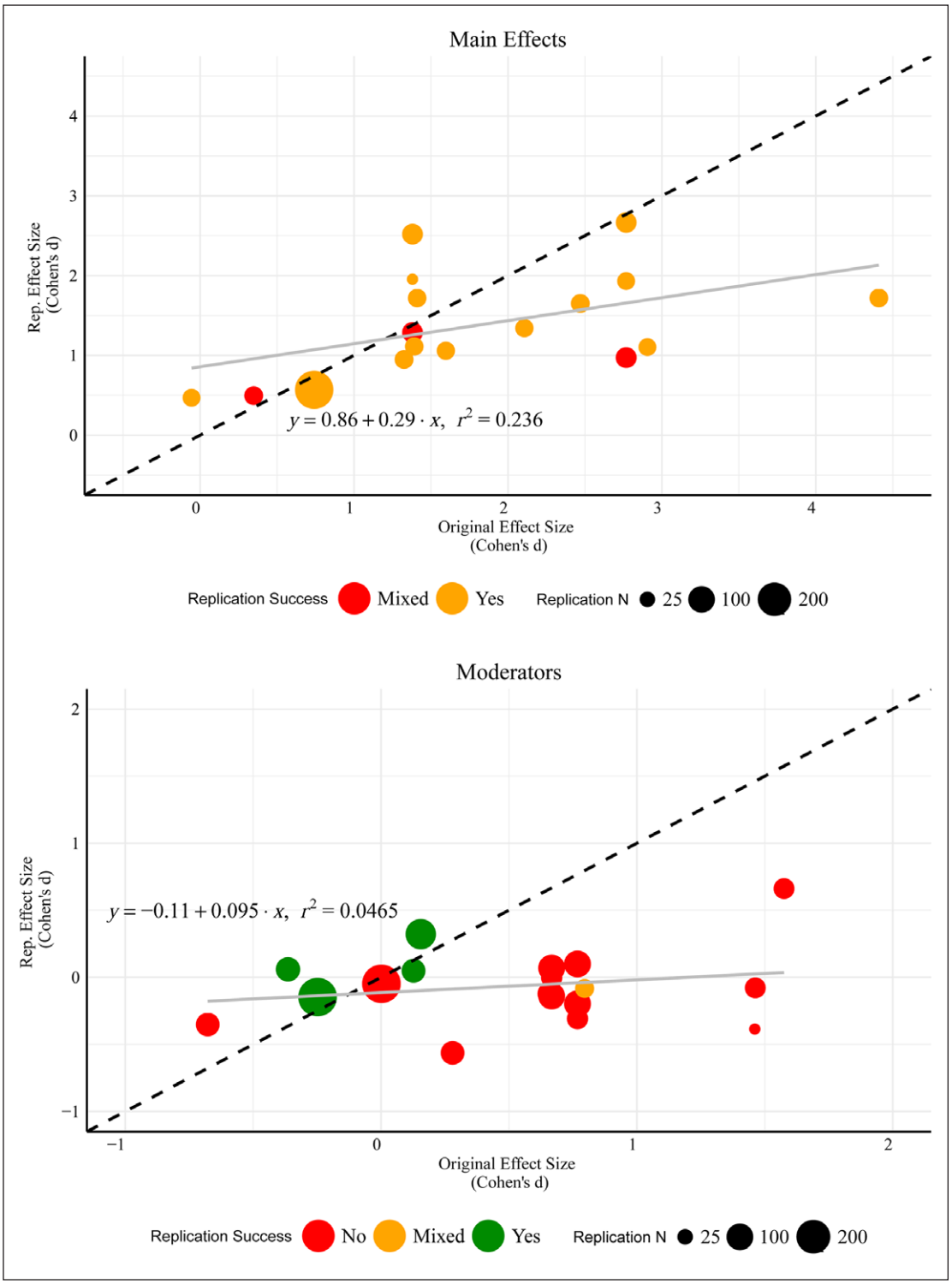

Figure 4: Scatterplot of replication effect size (Cohen's d) by original effect sizes. Note that the details provided in the original papers only allow us to calculate effect sizes based on percent correct (see main text for why this is problematic)—with the exception of Frank et al. (2010), for which the raw data are publicly available.

Size represents N of the replication comparison and color represents replication results. Perfect replication would result in a regression line through the origin with a slope of 1 (dotted line). A complete failure to replicate would result in a slope of 0. The actual slope for main effects (top panel) is .29, indicating that effects tend to be smaller in the replications than in the original (a steeper slope would have meant the opposite). For the moderators, there is no relationship between the original effect size and the replication effect size (B = 0.095).

For the original results effect sizes were calculated as follows. For between-subject or single-sample tests, we used the t-statistic (if available). For Frank et al. (2010), we derived Cohen's d from parameter estimates for the continuous-sentence-length model. For the remaining, we calculated Cohen's d using means and pooled variance. In many of these cases, no information about variance was reported, in which case we used the variance from our own data. For the replications, we calculated effect sizes after first excluding subjects with poor catch trial accuracy.

Table 2: Effect size for main effect of statistical word segmentation.

		Original	Replication	Included in Figure 4
SNA96-1a	part-word condition – chance	15**	12*	yes
	Nonword condition – chance	26**	15***	yes
	bupabi – chance	24***	57	no
SNA96-1b	part-word condition – chance	15**	18***	yes
	Nonword condition – chance	26**	20***	yes
	bupabi	24***	71***	no
SNA96-1c	part-word condition – chance	15**	8	yes
	Nonword condition – chance	26**	7	yes
SJAN99-1	Language One – chance	24*	17***	yes
	Language Two – chance	30*	18***	yes
SJAN99-2	Language One – chance	14*	11***	yes
	Language Three – chance	16*	12***	yes
FHK08-1	Nonword control – chance	19*	21***	yes
	Nonword experimental – chance	25*	12*	yes
	Split-word control – chance	14*	12**	yes
	Split-word experimental – chance	–1	6	yes
FHK08-3	Nonword – chance	18*	12*	yes
	Split-word – chance	5	8	yes
FGGT10	part-word – chance	29*	24*	yes

Note: Asterisks represent significant difference (*$p < .05$,**$p < .01$,***$p < .001$) and are based on the revised analyses (binomial mixed effects models) for the replications. Effect sizes are differences in percent correct. Data here reflect results after excluding subjects with poor catch trial accuracy, except for SNA96-1c where there were no catch trials. In all but one case, we calculated main effects by investigating the mean of subject means. For Frank et al. (2010), we never calculated these means, so we estimated the main effects by applying the reverse logit function on the intercept for the continuous-sentence-length model.

and the original. However, patterns across replications can be more interpretable, and we see a clear pattern of robust replication of statistical learning itself but not of the moderators. This pattern persisted even when using a university subject pool (Exp. 1b), when using the same stimuli as the original (the two Finn & Hudson Kam replications), and whether or not inattentive subjects were excluded (Exps. 1a & 1b). Thus, explaining away these failures to replicate would likely require different explanations for each.

Another possibility is that the failures to replicate are due to the fact that we replicated the experiments *as reported* and not necessarily *as originally conducted*. It is possible that the non-replicated findings depended on features of the experiment that were not reported and may be unknown to the original experimenters themselves. These are hardly comforting thoughts: both imply that the literature is misleading, and the latter raises doubts as to whether any meaningful science is possible. Note that this line of argumentation again requires a different explanation for each failure to replicate.

In contrast, the entire pattern of results is consistent with what might be expected based on what we know of typical statistical power in psychology. Most of the replicated findings involved single-sample tests, whereas nearly all the non-replicated findings involved two-sample tests or interactions, which require more power. For the most part, the original studies had only very limited power. For instance, many of the experiments above had only 12 subjects per condition for between-subject comparisons, which is only enough power to detect a relatively large effect of condition (Cohen's $d = 0.8$) around 47% of the time. Counter-intuitively, low power is associated with false positives as well as false negatives (Bakker et al., 2012; Button et al., 2013; Hartshorne & Schachner, 2012). Notably, the one experiment that replicated most fully — Frank et al. (2010) — involved an order of magnitude more subjects than the other five. Moreover, the one finding from this experiment that did *not* replicate was a surprising null effect of word length. The fact that this effect was significant and in the expected direction in our even more highly powered replication is consistent with power issues.

Thus, at the very least the findings that we were unable to replicate should be treated with caution absent additional evidence. Moreover, it may be sensible to

Table 3: Effect size for moderators of statistical word segmentation.

		Original	Replication	Included in Figure 4
SNA96-1a	nonword – part-word	11*	3	yes
	High TP – low TP	7*	0	yes
	Change-first – change-final foils	19**	5	yes
SNA96-1b	nonword – part-word	11*	2	yes
	High TP – low TP	7*	2	yes
	Change-first – change-final foils (false alarm rates)	19**	2	yes
SNA96-1c	nonword – part-word	11*	−1	yes
	High TP – low TP	7*	−1	yes
	Change-first – change-final foils (false alarm rates)	19**	−8	yes
SJAN99-1	Language One – Language Two	−6	−1	yes
	Lang 1: high TP – low TP	(*unreported*)*	4	no
	Lang 2: high TP – low TP	(*unreported*)*	0	no
SJAN99-2	Language One – Language Three	−2	−1	yes
	Change-first – change-final foils	(*unreported*)*	2	yes
	Exp. 2 – Exp. 1	3	−5*	yes
FHK08-1	nonword: control – experimental	−6	9	yes
	Split-word: control – experimental	15*	6	yes
FHK08-3	nonword – split-word	13*	4	yes
	Kmodu – nonword	32*	21	no
	Kmodu – split-word	45*	25	no
FGGT10	Effect of word length (B)	0.00	.11***	yes
	Effect of sentence length (B)	−0.06*	0.04***	yes

Note: Asterisks represent significant difference (*p < .05,**p < .01,***p < .001) and are based on the revised analyses (binomial mixed effects models) for the replications. Effect sizes are either differences in percent correct or (for Frank et al., 2010) regression coefficients. All effects are calculated after excluding subjects with poor catch trial accuracy, except for SNA96-1c where there were no catch trials.

apply similar caution to other findings with regards to moderators of adult statistical word segmentation, unless the experiment in question had unusually high power.

It should be noted that some of these effects are supported by additional findings in the literature, mostly in the form of testing the same question with minor methodological variations ("parametric extensions"). However, in the present study, we not only failed to replicate the original reports for (1), (2), and (3), but also parametric extensions of the same. Thus, we reserve judgment on the import of any additional conceptual replications in the literature until these, too, have been replicated.

Broader Implications

The present results raise concerns about the reliability of many of the findings in the adult statistical word segmentation literature. This worry must be preliminary, given that we only investigated six experiments chosen semi-randomly from among the more highly-cited papers in the literature. However, it is consistant both with broad-based investigations of the psychological literature (Open Science Collaboration, 2015; Vankov et al., 2014) and a recent meta-analysis of the infant statistical word segmentation literature, which revealed chronically low power (Black & Bergmann, 2017).

While the findings of this project are most straightforwardly applicable to the statistical word segmentation literature, there are broader implications. As already mentioned, opinions vary widely as to the degree to which psychological science is robust (Button et al., 2013; Gilbert et al., 2016; Ioannidis, 2012). The language sciences have many of the features that are believed to be protective: effects are often large, research is theory-driven, data-sharing and stimulus-sharing are common, and papers often involve multiple replications or parametric extensions of the key findings (Ferreira & Henderson, 2017). While we generally agree with this line of argumentation, the present results give pause, since the papers investigated contained many of those protective features.

We end on a methodological point. Many investigations of replicability and reproducibility aim for broad coverage. There are obvious advantages to that, but it means mastering many different paradigms and methods. For instance, the Open Science Collaboration (2015) replicated 100 different experiments, requiring 270 authors and many years to do so. By focusing on a narrowly-defined literature — which has theoretical advantages as well (see Introduction) — we were able to reuse the same experimental pipeline for all eight experiments. Having worked on both projects, we can confirm that the present method was much easier, and we recommend it to others interested in determining which are the most robust findings in a literature of central importance to their research.

Data Accessibility Statement

All data, materials, and code are available at osf.io/ehu7q/.

Exp. 1b was preregistered at osf.io/k2hu4/.

Notes

[1] An ongoing meta-analysis has so far identified more than 100 experiments; it is unclear how many more there may be.

[2] Using a standardized sample size has several advantages over choosing sample size for each experiment based on a power analysis, as was done in some previous replication studies (Camerer et al., 2018; Open Science Collaboration, 2015): It facilitates comparison across experiemnts, which otherwise may have widely varying precision; we need not chose a single effect per experiment to focus on, and; it simplifies experiment creation. Moreover, prior replication studies have found that published effect sizes are often so imprecise as to significantly diminish the usefulness of power analyses.

[3] For instance, one option is to ask whether the effect size estimate in the replication is within the confidence interval for the original. If the original has low power, however, that confidence interval may be so wide as to make replication trivial.

[4] This reflects a shift in our method from collecting 50 subjects per condition prior to the attention screen, which we used in our initial experiments, to collecting 50 subjects per condition exclusive of excluded subjects. Though reported at the beginning, Exp. 1b was the last experiment in this paper to be run.

[5] This reflects the fact that Exp. 1c was actually the first to be run. We decided to add catch trials to subsequent experiments after reviewing Exp. 1c's results.

[6] We ran a mixed effects binomial regression with a fixed effect of setting (in-lab vs. online) and random intercept of word/foil pair. Note that the model did not converge with the random intercept of subject, so it was removed from the model.

[7] We had not yet considered the possibility of putting all catch trials at the end. Though reported last, this experiment was the first to be run with catch trials.

[8] This finding must be somewhat tempered by the fact that the more complex model failed to fully converge, which is itself further reason to not investigate it further.

[9] The one exception is the AMT+WatsonTTS replication of Saffran et al. (1996) Exp. 1. That failure could be explained a number of ways, the most parsimonius being random chance: the probability of observing an effect nine consecuitive times is low, even when the effect is real (Francis, 2012).

Acknowledgements

We thank Elissa Newport, Richard Aslin, Jenny Saffran, Mike Frank, Roman Feiman, Charlie Ebersole, Hugh Rabagliati, Miguel Mejia, Hayley Greenough, and three anonymous reviewers for insightful comments and feedback, and to Amy Finn for providing assistance in replicating her work.

Funding Information

This work was partially supported by a TAME grant from Boston College.

Competing Interests

The authors have no competing interests to declare.

Author Contributions

All authors contributed to design, acquisition, analysis, interpretation, and drafting. Final drafting was done by JKH and LS. Final approval: JKH.

References

Aarts, A. A., & **LeBel, E. P.** (2016). Curate science: A platform to gauge the replicability of psychological science.

Ambridge, B., Pine, J. M., Rowland, C. F., Jones, R. L., & **Clark, V.** (2009). A semantics-based approach to the "no negative evidence problem". *Cognitive Science*, *33*(7), 1301–1316. DOI: https://doi.org/10.1111/j.1551-6709.2009.01055.x

Anderson, C., Bahnik, S., Barnett-Cowan, M., Bosco, F., Chandler, J., Chartier, C., Zuni, K., et al. (2016). Response to comment on "estimating the reproducibility of psychological science". *Science*, *351*(62–77). DOI: https://doi.org/10.1126/science.aad9163

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., Wicherts, J. M., et al. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*(2), 108–119. DOI: https://doi.org/10.1002/per.1919

Baayen, R. H., Davidson, D. J., & **Bates, D. M.** (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. DOI: https://doi.org/10.1016/j.jml.2007.12.005

Bagwell, C., & **Contributors, S.** (2015). *Sox: Sound exchange, the swiss army knife of audio manipulation*. sox.sourceforge.net.

Baker, C. L. (1979). Syntactic theory and the projection problem. *Linguistic Inquiry*, *10*(4), 533–581.

Bakker, M., Hartgerink, C. H., Wicherts, J. M., & **van der Maas, H. L.** (2016). Researchers' intuitions

about power in psychological research. *Psychological Science*, *27*(8), 1069–1077. DOI: https://doi.org/10.1177/0956797616647519

Bakker, M., van Dijk, A., & **Wicherts, J. M.** (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543–554. DOI: https://doi.org/10.1177/1745691612459060

Berinsky, A. J., Huber, G. A., & **Lenz, G. S.** (2012). Evaluating online labor markets for experimental research: Amazon.com's mechanical turk. *Political Analysis*, *20*(3), 351–368. DOI: https://doi.org/10.1093/pan/mpr057

Bezeau, S., & **Graves, R.** (2001). Statistical power and effect sizes of clinical neuropsychology research. *Journal of Clinical and Experimental Neuropsychology*, *23*(3), 399–406. DOI: https://doi.org/10.1076/jcen.23.3.399.1181

Black, A., & **Bergmann, C.** (2017). Quantifying infants' statistical word segmentation: A meta-analysis. In *Proceedings of the 39th annual conference of the cognitive science society*, 124–129.

Bonatti, L. L., Peña, M., Nespor, M., & **Mehler, J.** (2005, Jun). Linguistic constraints on statistical computations: The role of consonants and vowels in continuous speech processing. *Psychol Sci*, *16*(6), 451–9.

Bowerman, M. (1988). The 'no negative evidence' problem: How do children avoid constructing an overly general grammar. In: Hawkins, J. A. (Ed.), *Explaining language universals*, 73–101. Malden, MA: Blackwell.

Brown, R., & **Hanlon, C.** (1970). Derivational complexity and order of acquisition in child speech. In: Hayes, J. R. (Ed.), *Cognition and the development of language*. New York: Willy.

Buhrmester, M., Kwang, T., & **Gosling, S. D.** (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5. DOI: https://doi.org/10.1177/1745691610393980

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*(5), 365–376. DOI: https://doi.org/10.1038/nrn3475

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433–1436. DOI: https://doi.org/10.1126/science.aaf0918

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., et al. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, 1. DOI: https://doi.org/10.1038/s41562-018-0399-z

Chandler, J., Mueller, P., & **Paolacci, G.** (2014). Nonnaïveté among amazon mechanical turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, *46*(1), 112–130. DOI: https://doi.org/10.3758/s13428-013-0365-7

Chase, L. J., & **Chase, R. B.** (1976). A statistical power analysis of applied psychological research. *Journal of Applied Psychology*, *61*(2), 234–237. DOI: https://doi.org/10.1037/0021-9010.61.2.234

Chemla, E., Mintz, T. H., Bernal, S., & **Christophe, A.** (2009). Categorizing words using 'frequent frames': What cross-linguistic analyses reveal about distributional acquisition strategies. *Developmental Science*, *12*(3), 396–406. DOI: https://doi.org/10.1111/j.1467-7687.2009.00825.x

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*(4), 335–359. DOI: https://doi.org/10.1016/S0022-5371(73)80014-3

Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., Attie, M., et al. (2018). Estimating the reproducibility of experimental philosophy. *PsyArXiv*, *21*. DOI: https://doi.org/10.1007/s13164-018-0400-9

Crane, H. (2017). Why redefining statistical significance will not improve reproducibility and could make the replication crisis worse. *SSRN*. DOI: https://doi.org/10.2139/ssrn.3074083

Crump, M. J. C., McDonnell, J. V., & **Gureckis, T. M.** (2013). Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PLoS ONE*, *8*(3). DOI: https://doi.org/10.1371/journal.pone.0057410

De Leeuw, J. R. (2015). jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior research methods*, *47*(1), 1–12. DOI: https://doi.org/10.3758/s13428-014-0458-y

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., & **Van der Vrecken, O.** (1996). The mbrola project: Towards a set of high quality speech synthesizers free of use for non commercial purposes. In: *Proceedings of the Fourth International Conference on Spoken Language*, *3*, 1393–1396. DOI: https://doi.org/10.1109/ICSLP.1996.607874

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Nosek, B. A., et al. (2016). Many labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68–82. DOI: https://doi.org/10.1016/j.jesp.2015.10.012

Fanelli, D. (2009). How many scientists fabricate and falsify research? a systematic review and meta-analysis of survey data. *PloS one*, *4*(5), e5738. DOI: https://doi.org/10.1371/journal.pone.0005738

Fernandes, T., Ventura, P., & **Kolinsky, R.** (2011). The relative weight of statistical and prosodic cues in speech segmentation: A matter of language-(in) dependency and of signal quality. *Journal of Portuguese Linguistics*, *10*(1). DOI: https://doi.org/10.5334/jpl.102

Ferreira, F., & **Henderson, J. M.** (2017). *Defending .05: It's not enough to be suggestive* (Blog No. July 26).

rolfzwaan.blogspot.nl/2017/07/defending-05-its-not-enough-to-be_26.html.

Finn, A. S., & **Hudson Kam, C. L.** (2008). The curse of knowledge: First language knowledge impairs adult learners' use of novel statistics for word segmentation. *Cognition*, *108*(2), 477–499. DOI: https://doi.org/10.1016/j.cognition.2008.04.002

Fraley, R. C., & **Vazire, S.** (2014). The n-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PloS one*, *9*(10), e109019. DOI: https://doi.org/10.1371/journal.pone.0109019

Francis, G. (2012). The Psychology of Replication and Replication in Psychology. *Perspectives on Psychological Science*, *7*(6), 585–594. DOI: https://doi.org/10.1177/1745691612459520

Franco, A., Cleeremans, A., & **Destrebecqz, A.** (2011). Statistical learning of two artificial languages presented successively: How conscious? *Frontiers in psychology*, *2*, 229. DOI: https://doi.org/10.3389/fpsyg.2011.00229

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., Yurovsky, D., et al. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy*, *22*(4), 421–435. DOI: https://doi.org/10.1111/infa.12182

Frank, M. C., Goldwater, S., Griffiths, T. L., & **Tenenbaum, J. B.** (2010). Modeling human performance in statistical word segmentation. *Cognition*, *117*(2), 107–125. DOI: https://doi.org/10.1016/j.cognition.2010.07.005

Frank, M. C., Lewis, M. L., & **MacDonald, K.** (2016). A performance model for early word learning. In: *Proceedings of the 38th annual conference of the cognitive science society*.

Garcia, C. R., Iozzo, G. L., Lamirato, K. E., Ledoux, J. R., Mu, J., Murdock, K. N., Hartshorne, J. K., et al. (2017). Replication of Saffran, Newport, & Aslin (1996) Word segmentation: The role of distributional cues, exp. 1. DOI: https://doi.org/10.31234/osf.io/qsyd2

Garcia, C. R., van Horne, K. D., & **Hartshorne, J. K.** (2017). Replication of finn & hudson kam (2008) the curse of knowledge: First language knowledge impairs adult learners' use of novel statistics for word segmentation, exp. 1. DOI: https://doi.org/10.31234/osf.io/2xcwk

Gibson, E., & **Fedorenko, E.** (2010). Weak quantitative standards in linguistics research. Trends in *Cognitive Sciences*, *14*(6), 233. DOI: https://doi.org/10.1016/j.tics.2010.03.005

Gibson, E., & **Fedorenko, E.** (2011). The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes*, *28*(1), 88–124.

Gilbert, D., King, G., Pettigrew, S., & **Wilson, T.** (2016). More on "Estimating the reproducibility of psychological science". *Science*, *351*(6277), 1037. DOI: https://doi.org/10.1126/science.aad7243

Goldwater, S., Griffiths, T. L., & **Johnson, M.** (2009). A bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*(1), 21–54. DOI: https://doi.org/10.1016/j.cognition.2009.03.008

Gray, K., & **Wegner, D. M.** (2013). Six guidelines for interesting research. *Perspectives on Psychological Science*, *8*(5), 549–553. DOI: https://doi.org/10.1177/1745691613497967

Hardwicke, T. E., Mathur, M., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., et al. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal cognition.

Hartshorne, J. K. (2017). Replication of saffran, newport, & aslin (1996) word segmentation: The role of distributional cues, exp. 1. DOI: https://doi.org/10.31234/osf.io/e5c64

Hartshorne, J. K., & **Schachner, A.** (2012). Tracking replicability as a method of post-publication open evaluation. *Frontiers in computational neuroscience*, *6*(8). DOI: https://doi.org/10.3389/fncom.2012.00008

Hartshorne, J. K., & **Skorb, L.** (2018). In-lab replication of saffran, newport, & aslin (1996) word segmentation: The role of distributional cues, exp. 1.

Hauser, D. J., & **Schwarz, N.** (2016). Attentive turkers: Mturk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, *48*(1), 400–407. DOI: https://doi.org/10.3758/s13428-015-0578-z

Henrich, J., Heine, S. J., & **Norenzayan, A.** (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. DOI: https://doi.org/10.1017/S0140525X0999152X

Hoch, L., Tyler, M. D., & **Tillmann, B.** (2013). Regularity of unit length boosts statistical learning in verbal and nonverbal artificial languages. *Psychonomic bulletin & review*, *20*(1), 142–147. DOI: https://doi.org/10.3758/s13423-012-0309-8

IBM. (2017). *Watson developer cloud*. www.ibm.com/watson/developercloud/text-to-speech.html.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. DOI: https://doi.org/10.1371/journal.pmed.0020124

Ioannidis, J. P. A. (2012). Why Science Is Not Necessarily Self-Correcting. *Perspectives on Psychological Science,* *7*(6), 645–654. DOI: https://doi.org/10.1177/1745691612464056

Iozzo, G. I., Lamirato, K. E., & **Hartshorne, J. K.** (2017). Replication of finn & hudson kam (2008) the curse of knowledge: First language knowledge impairs adult learners' use of novel statistics for word segmentation, exp. 3.

Ipeirotis, P. G. (2010). Demographics of mechanical turk. *NYU Working Papers*(CEDER-10-01).

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*,

59(4), 434–446. DOI: https://doi.org/10.1016/j.jml.2007.11.007

Johnson, E. K., & **Tyler, M. D.** (2010). Testing the limits of statistical learning for word segmentation. *Developmental science*, *13*(2), 339–345. DOI: https://doi.org/10.1111/j.1467-7687.2009.00886.x

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality*, *39*(1), 103–129. DOI: https://doi.org/10.1016/j.jrp.2004.09.009

Judd, C. M., Westfall, J., & **Kenny, D. A.** (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of personality and social psychology*, *103*(1), 54–69. DOI: https://doi.org/10.1037/a0028347

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196–217. DOI: https://doi.org/10.1207/s15327957pspr0203_4

Klatzky, R., Au, T., Bar, M., Berntsen, D., Dawson, G., Hatfield, E. A., Weber, E., et al. (2018). Firm foundations: Leading researchers name the most replicated findings in psychological science. *Observer*, *31*(1).

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., Nosek, B. A., et al. (2014). Investigating variation in replicability. *Social Psychology*. DOI: https://doi.org/10.1027/1864-9335/a000178

Kovacs, A. M., & **Mehler, J.** (2009). Cognitive gains in 7-month-old bilingual infants. *Proceedings of the National Academy of Sciences*, *106*(16), 6556–6560. DOI: https://doi.org/10.1073/pnas.0811323106

Kurumada, C., Meylan, S. C., & **Frank, M. C.** (2013). Zipfian frequency distributions facilitate word segmentation in context. *Cognition*, *127*(3), 439–453. DOI: https://doi.org/10.1016/j.cognition.2013.02.002

Lewis, M., Braginsky, M., Tsuji, S., Bergmann, C., Piccinini, P., Cristia, A., & **Frank, M. C.** (2016). A quantitative synthesis of early language acquisition using meta-analysis.

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive therapy and research*, *1*(2), 161–175. DOI: https://doi.org/10.1007/BF01173636

Mahowald, K., James, A., Futrell, R., & **Gibson, E.** (2017). A meta-analysis of syntactic priming in language production. *Journal of Memory and Language*.

Makel, M. C., Plucker, J. A., & **Hegarty, B.** (2012). Replications in Psychology Research: How Often Do They Really Occur? *Perspectives on Psychological Science*, *7*(6), 537–542. DOI: https://doi.org/10.1177/1745691612460688

Marcus, G. F. (1993). Negative evidence in language acquisition. *Cognition*, *46*(1), 53–85. DOI: https://doi.org/10.1016/0010-0277(93)90022-N

Marcus, G. F., & **Berent, I.** (2003). Are there limits to statistical learning? *Science*, *300*(5616), 53–55. DOI: https://doi.org/10.1126/science.300.5616.53

Mason, W., & **Suri, S.** (2011). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, *44*(1), 1–23. DOI: https://doi.org/10.3758/s13428-011-0124-6

Maxwell, S. E., Lau, M. Y., & **Howard, G. S.** (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, *70*(6), 487–498. DOI: https://doi.org/10.1037/a0039400

Mone, M. A., Mueller, G. C., & **Mauland, W.** (1996). the Perceptions and Usage of Statistical Power in Applied Psychology and Management Research. *Personnel Psychology*, 103–120. DOI: https://doi.org/10.1111/j.1744-6570.1996.tb01793.x

Morey, R. D., & **Lakens, D.** (2016). *Why most of psychology is statistically unfalsifiable*. Submitted.

Mu, J., Ledoux, J., & **Hartshorne, J.** (2017). Replication of frank, goldwater, griffiths, & tenenbaum (2010): Modeling human performance in statistical word segmentation, experiment 1.

Murdock, K. N., Savery, A. A., Trimm, K. A., Vidal, J., & **Hartshorne, J. K.** (2017a). Replication of saffran, johnson, aslin, & newport (1999) statistical learning of tone sequences by human infants and adults, exp. 1.

Murdock, K. N., Savery, A. A., Trimm, K. A., Vidal, J., & **Hartshorne, J. K.** (2017b). Replication of saffran, johnson, aslin, & newport (1999) statistical learning of tone sequences by human infants and adults, exp. 2.

Newton, D. P. (2010). Quality and peer review of research: An adjudicating role for editors. *Accountability in Research*, *17*(3), 130–145. DOI: https://doi.org/10.1080/08989621003791945

Nosek, B. A., & **Lakens, D.** (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology*, *45*(3), 137–141. DOI: https://doi.org/10.1027/1864-9335/a000192

Nosek, B. A., Spies, J. R., & **Motyl, M.** (2012). Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability. *Perspectives on Psychological Science*, *7*(6), 615–631. DOI: https://doi.org/10.1177/1745691612459058

Olson, C. M., Rennie, D., Cook, D., Dickersin, K., Flanagin, A., Hogan, J. W., Pace, B., et al. (2002). Publication bias in editorial decision making. *JAMA: The journal of the American Medical Association*, *287*(21), 2825–2828. DOI: https://doi.org/10.1001/jama.287.21.2825

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*. DOI: https://doi.org/10.1126/science.aac4716

Paolacci, G., Chandler, J., & **Ipeirotis, P. G.** (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*(5), 411–419.

Pashler, H., & **Harris, C. R.** (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, *7*(6), 531–536. DOI: https://doi.org/10.1177/1745691612463401

Pearl, L., Goldwater, S., & **Steyvers, M.** (2010). Online learning mechanisms for Bayesian models of word segmentation. *Research on Language and Computation*, *8*(2–3), 107–132. DOI: https://doi.org/10.1007/s11168-011-9074-5

Perfors, A., Tenenbaum, J. B., & **Wonnacott, E.** (2010). Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of Child Language*, *37*(3), 607–642. DOI: https://doi.org/10.1017/S0305000910000012

Perone, M. (2018). How i learned to stop worrying and love replication failures. *Perspectives on Behavior Science*, 1–18. DOI: https://doi.org/10.1007/s40614-018-0153-x

Perruchet, P., & **Desaulty, S.** (2008). A role for backward transitional probabilities in word segmentation? *Memory & Cognition*, *36*(7), 1299–1305. DOI: https://doi.org/10.3758/MC.36.7.1299

Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.

Quine, W. V. O. (1960). *Word and object*. Cambridge, UK: Cambridge University Press.

Rand, D. G. (2012). The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology*, *299*, 172–179. DOI: https://doi.org/10.1016/j.jtbi.2011.03.004

Rhoades, L. J. (2004). Ori closed investigations into misconduct allegations involving research supported by the public health service: 1994–2003. Investigations1994-2003-2.pdf.

Richard, F. D., Bond, C. F., Jr., & **Stokes-Zoota, J. J.** (2003). One Hundred Years of Social Psychology Quantitatively Described. *Review of General Psychology*, *7*(4), 331–363. DOI: https://doi.org/10.1037/1089-2680.7.4.331

Romberg, A. R., & **Saffran, J. R.** (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(6), 906–914. DOI: https://doi.org/10.1002/wcs.78

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological bulletin*, *86*(3), 638. DOI: https://doi.org/10.1037/0033-2909.86.3.638

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of consulting and clinical psychology*, *58*(5), 646. DOI: https://doi.org/10.1037/0022-006X.58.5.646

Rouse, S. V. (2015). A reliability analysis of mechanical turk data. *Computers in Human Behavior*, *43*, 304–307. DOI: https://doi.org/10.1016/j.chb.2014.11.004

Saffran, J. R., Aslin, R. N., & **Newport, E. L.** (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928. DOI: https://doi.org/10.1126/science.274.5294.1926

Saffran, J. R., Johnson, E. K., Aslin, R. N., & **Newport, E. L.** (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*(1), 27–52. DOI: https://doi.org/10.1016/S0010-0277(98)00075-4

Saffran, J. R., Newport, E. L., & **Aslin, R. N.** (1996). Word segmentation: The role of distributional cues. *Journal of memory and language*, *35*(4), 606–621. DOI: https://doi.org/10.1006/jmla.1996.0032

Schweinsberg, M., Madan, N., Vianello, M., Sommer, S. A., Jordan, J., Tierney, W., et al. (2016). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology*, *66*, 55–67. DOI: https://doi.org/10.1016/j.jesp.2015.10.001

Shapiro, D. N., Chandler, J., & **Mueller, P. A.** (2013). Using mechanical turk to study clinical populations. *Clinical Psychological Science*, *1*(2), 213–220. DOI: https://doi.org/10.1177/2167702612469015

Shen, W., Kiger, T. B., Davies, S. E., Rasch, R. L., Simon, K. M., & **Ones, D. S.** (2011). Samples in applied psychology: Over a decade of research in review. *Journal of Applied Psychology*, *96*(5), 1055–1064. DOI: https://doi.org/10.1037/a0023322

Simons, D. J., Holcombe, A. O., & **Spellman, B. A.** (2014). An introduction to registered replication reports at perspectives on psychological science. *Perspectives on Psychological Science*, *9*(5), 552–555. DOI: https://doi.org/10.1177/1745691614543974

Spellman, B. A. (2012). Introduction to the special section: Data, data, everywhere… especially in my file drawer. *Perspectives on Psychological Science*, *7*(1), 58–59. DOI: https://doi.org/10.1177/1745691611432124

Stroebe, W. (2016). Are most published social psychological findings false? *Journal of Experimental Social Psychology*, *66*, 134–144. DOI: https://doi.org/10.1016/j.jesp.2015.09.017

Stroebe, W., Postmes, T., & **Spears, R.** (2012). Scientific Misconduct and the Myth of Self-Correction in Science. *Perspectives on Psychological Science*, *7*(6), 670–688. DOI: https://doi.org/10.1177/1745691612460687

Thiessen, E. D. (2017). What's statistical about learning? insights from modelling statistical learning as a set of memory processes. *Phil. Trans. R. Soc. B*, *372*(1711). DOI: https://doi.org/10.1098/rstb.2016.0056

Tomasello, M. (2009). *Constructing a language*. Harvard university press.

Toro, J. M., Sinnett, S., & **Soto-Faraco, S.** (2005). Speech segmentation by statistical learning depends on attention. *Cognition*, *97*(2), B25–B34. DOI: https://doi.org/10.1016/j.cognition.2005.01.006

Vankov, I., Bowers, J., & **Munafò, M. R.** (2014). On the persistence of low power in psychological science. *The Quarterly Journal of Experimental Psychology*, *67*(5), 1037–1040. DOI: https://doi.org/10.1080/17470218.2014.885986

Vul, E., Harris, C., Winkielman, P., & **Pashler, H.** (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, *4*(3), 274–290. DOI: https://doi.org/10.1111/j.1745-6924.2009.01125.x

Zwaan, R. A., Etz, A., Lucas, R. E., & **Donnellan, M. B.** (n.d.). Making replication mainstream. *Behavioral and Brain Sciences*, *40*, 1–50.

How to cite this article: Hartshorne, J. K., Skorb, L., Dietz, S. L., Garcia, C. R., Iozzo, G. L., Lamirato, K. E., Ledoux, J. R., Mu, J., Murdock, K. N., Ravid, J., Savery, A. A., Spizzirro, J. E., Trimm, K. A., van Horne, K. D., & Vidal, J. (2019). The Meta-Science of Adult Statistical Word Segmentation: Part 1. *Collabra: Psychology,* 5(1): 1. DOI: https://doi.org/10.1525/collabra.181

Senior Editor: Rolf Zwaan

Editor: Max Coltheart

Submitted: 06 July 2018 **Accepted:** 01 October 2018 **Published:** 08 January 2019

UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Collabra: Psychology is a peer-reviewed open access journal published by University of California Press.

OPEN ACCESS