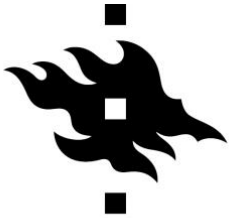

UNIVERSITY OF HELSINKI

Replication science in theory and practice:

Two replications of a quadrisyllabic word segmentation study

Master's Program in Linguistic Diversity and Digital Humanities, Cognitive Science

Master's thesis

Author:

Nicholas Nelson

Supervisors:

Dr. Soila Kuuluvainen

Professor Riikka Möttönen

20.5.2025

Helsinki

Title: Replication science in theory and practice: Two replications of a quadrisyllabic word segmentation study

Author: Nicholas Nelson

Month and year: May 2025

Number of pages: 67

Keywords: replication, replication crisis, statistical language learning, stimuli creation

Abstract: The aim of this thesis is to focus on replication science using statistical language learning as a case study in replication completed by students. Replication is foundational to the advancement of knowledge. However, a replication crisis has emerged across disciplines, in which larger numbers of published studies cannot be replicated. Part I explores this crisis and its potential causes and highlights current efforts to provide solutions. One suggested solution is utilizing students to produce replication work to learn about the scientific process and the replication crisis, all while producing valuable knowledge to their fields.

Part II attempts a case study of replication by using two student experiments—one conceptual and one direct replication—all which focus on quadrisyllabic word segmentation in statistical language learning. Previous research has suggested that participants during quadrisyllabic word segmentation tasks successfully learn words from an exposure stream only when aided by 25 millisecond pauses between words but not if it is continuous (Benjamin et al., 2023). The original study was completed on French participants, in which the average word length is shorter than Finnish. Experiment 1 is a Finnish-language conceptual replication of Benjamin et al. using two artificial languages to accommodate for Finnish vowel harmony. We hypothesized that since Finnish has longer words on average that their word-length expectations would be longer, and thus segment quadrisyllabic words without the need for a pause, while still reporting more familiarity when a pause is included. Results suggest that Finnish speakers may not need pauses to segment quadrisyllabic word formations from continuous speech streams. Nevertheless, pauses help with word-form extraction also in Finnish speakers. After a focus on stimuli creation, Experiment 2 attempts a direct replication of Experiment 1. When distributed into language groups, results show support for the original claim that quadrisyllabic word segmentation is successful when aided by a 25 millisecond pause but not in a continuous stream. However, the combined results from both languages suggest support for our word-length hypothesis.

Part III discusses the vowel-placement differences, which may indicate that those sounds are processed differently during quadrisyllabic word segmentation paradigms; resultingly,

controlling for stimuli and its features is increasingly important. Finally, it is suggested that students can contribute to a viable solution to the replication crisis.

Table of contents

1 Part I: Replication: Theory and background

Replication has been described as the "cornerstone of science" (Simons, 2014), the "most important criterion of scientific knowledge" (Rosenthal & Rosnow, 2008), and the "scientific gold standard" (Jasny et al., 2011). Even though some argue against the severity of their claims (Haig, 2022), it is safe to say that replication studies can act as a stopgap for questionable practices in science and aims to establish the current state of replicability for a discipline (Balafoutas et al., 2025). Since the beginning of the 21st century, an increasing concern has plagued academic disciplines: a substantial number of published studies do not replicate (Pashler & Wagenmakers, 2012; Open Science Collaboration, 2015). Academic communities are facing a replication/reproducibility crisis that is undermining scientific integrity. Some disciplines impacted are psychology (Open Science Collaboration, 2015; Stevens, 2017; Nosek et al., 2022), neuroscience (Button et al., 2013; Miłkowski et al., 2018), economics (Camerer et al., 2016; Ferraro & Shukla, 2023; Huntington-Klein et al., 2021; Miguel, 2021), education (Frias-Navarro et al., 2020), physics (Franklin, 2018) and medicine (Prinz et al., 2011; Gabelica et al., 2022), among others. This widespread phenomenon in science is debated and reviewed across disciplines, and many scientists offer their own solutions to alleviate the crisis. To understand the solutions, we must first explore the problem.

1.1 Theory

1.1.1 Theoretical foundations

What precisely does replication mean? Simons claims that replication supports science using a simple premise: *"if an effect is real and robust, any competent researcher should be able to obtain it when using the same procedures with adequate statistical power"* (Simons, 2014, p. 76). Replication as scientific foundation asserts that true findings are only maintained through continued testing. This continued testing is the reason science requires replication: if an effect is true, then repeated testing of that effect would also be true both within the original data and in newly collected data. Approaches to retesting are nuanced and each replication produces different kinds of knowledge.

1.1.2 Types of replications

Following the recommendation of the National Academies of Sciences, *reproducibility* (also called *computational reproducibility*) is the ability to produce consistent results using the same input data and computational methods, which may include codes, types of analyses, steps taken during the process, etc., and *replicability* is obtained through consistent results across studies attempting to answer the same research question (2019). The former may be understood as verifying the results of a previous experiment, whereas the latter is focused on generating new data to test the effect's breadth of generalizability.

Attempts at distinguishing replication into different types began as early as 1968, when Lykken parsed between *literal*, *operational*, and *constructive* replication. In 2009, Stefan Schmidt collected definitions of replication from Lykken (1968), Sargent (1981), Keppel (1982), and Hendrick (1991), and divided them to two classes. *Direct replication* (sometimes called *exact replication*) recreates conditions and procedures exact or as similar as possible to the original study (Schmidt, 2009; Diener & Biswas-Diener, 2025). "As similar as possible" prevails as a definition because it is argued that *exact* conditions is a temporal impossibility (Nosek et al., 2022; Shadish et al., 2002). *Conceptual replication* is when a study utilizes different methodologies to test the same research question (Diener & Biswas-Diener, 2025). Conceptual replication tests the generalizability of findings. Improving generalizability through replication may look like testing on a new population or in a new domain. For example, Saffran et al. (1996) first published a study in statistical language learning using infants and in the same year produced a conceptual replication using adults (Saffran, Newport, et al., 1996). In this way, conceptual replication produces exploratory research by testing the boundaries of an effect, while direct replication tests whether original results measure a true effect.

Each type of replication is important to the advancement of knowledge since each approach provides new information—either supporting or refuting the research question investigated (see Simons, 2014 for a commentary on direct replication as the best evidence for effects). Using these definitions as foundation, the following section will address the replication crisis.

1.2 The replication crisis

The Open Science Collaboration has shown that across some leading journals in psychology, direct replication was achieved in an average of 36% of the investigated studies (2015). This suggests that if someone was to select a finding at random within the selected journals for this study, there would be a 36% chance that the study would be replicable. Although it is alarmingly low, this number does not necessarily need to be 100%. The purpose of science is to test and retest new and innovative theories and studies; failures exist as new methods are tested and improved. The possibility of false positives in data remains a possibility as well (see 1.2.1 for more on Type I errors). The replication crisis encompasses both direct and conceptual replications, but what does this mean? Direct replications measure a true effect outlined by an original study. This could be non-replicable, especially if the measured effect includes qualitative measures that may be influenced by societal changes over time. However, if a failed direct replication cannot be explained by these qualitative measures, then some aspect of the original study should be investigated as to the potential reasons of non-replicability. Conceptual replication, on the other hand, produces exploratory research into an effect's generalizability. Non-replication in these studies merely discovers the boundaries of that generalizability or call for a reinterpretation of results. When discussing the replication crisis, most studies will focus on direct replication, as those produce the largest problem in replication. The low results from Open Science Collaboration indicate a warranted investigation into possible causes, potential results and effects, and ways the effect can be mitigated in further studies. This section will attempt to explore those areas: causes, effects on public trust, and current efforts and next steps.

1.2.1 Causes

It is suggested that the replication crisis is rooted in incentives within academic research and publishing (Smaldino & McElreath, 2016; Kohrt et al., 2023). Although replication is needed for the advancement of knowledge, completing replications has been faced with challenges. Some of those challenges are the suggested pressures that contribute to the replication crisis: the relationship between academic publishers and researchers and the individual practices by researchers.

1.2.1.1 Academic publishing and researchers

Publication bias is any behavior that favors manuscripts finding support for their tested hypotheses (Scheel et al., 2021). Publication bias has been discussed at length in the literature (Balafoutas et al., 2025), which suggests that unintended bias has influenced manuscript selection, not only in favoring studies that find support for tested hypotheses, but also a pressure toward novel studies and studies that have statistically significant results.

Within peer-reviewed publications (academic publishing) there has been and continues to be pressure towards novel studies (Stevens, 2017). This pressure often results in fewer replication studies being published. Even if a replication is published, they still retain some element of novelty. If we can conclude that conceptual replications are considered more novel than direct replications and we consider that conceptual replications are published more frequently than direct (Keating & Totzkay, 2019), then this would also support this publishing trend toward novelty. Not addressing novelty bias will retain or diminish the number of replication studies, since they are viewed as not valuable nor wanted for publication.

However, this bias towards novelty is not unique to publishers, editors, and reviewers; it is suggested to be a larger "disease" of research called *neophilia*: "an excessive appreciation for novelty" (Antonakis, 2017, p. 5). This could reasonably expand to any person impacted by the scientific process, including researchers, academics, research grant providers, among others.

Antonakis also suggests another disease in research production: *significosis*: "an inordinate focus on statistically significant results" (2017, p. 5). Type I errors, false positive results that reject the null hypothesis when it is true within the population, are potentially biasing academic publishing and spurring the replication crisis. Controlling for Type I errors is used through significance (alpha, α) levels, typically set at 0.05. This indicates results are computed assuming there is only a 5% chance of receiving false positive results. However, if publishers still favor significosis and the results indeed are a false positive (5% chance), this indicates a potential publication bias that may favor

Type I errors through significosis, resulting in the publication of false-positive results. This also suggests that the true effects (true negative results) may be present in works unpublished or hiding in the "file drawer" (Rosenthal, 1979). This file drawer problem also indicates that researchers may not be willing to submit their completed works for publication, since they're astutely aware of the publication trend to favor positive results.

The ways neophilia and significosis impact researchers is through a systemic pressure often referred to as "publish or perish," the idea that in order to survive or flourish as a researcher, one must continually publish work. At the core of this phrase is the relationship between publisher and researchers, where one continually learns from the other. Researchers are aware of neophilia and significosis in academic publishing, potentially impacting individual lines of inquiry. One result may be that researchers only aim to complete novel studies, whether original work or conceptual replication, resulting in both publishers and researchers unintentionally utilizing neophilia. The continued pressure to publish has also been identified as a possible strain that may encourage researchers to participate in questionable research practices in order to get published (Balafoutas et al., 2025).

1.2.1.2 Individual practices by researchers

Some aspects of individual behaviors in research have been identified as potential contributors to the replication crisis, mostly the incentivized behavior of doctoring results in order to be published. Questionable research practices (QRPs) are the widely addressed behaviors that are not inherently bad behaviors in research, rather they may provide incomplete (and thus false) information of analysis, results, or conclusions. Common QRPs are *p*- or null-hacking (manipulating analysis to enable a desired effect), selective reporting (only reporting a portion of the completed study), and HARKing (hypothesizing after results are known) (Balafoutas et al., 2025).

The pressure for young researchers to engage in QRPs lies in significosis. It is shown that marginally statistically significant results of economics PhD graduates have higher likelihoods of academic placement, and the hiring decisions of those academic

institutions show publication bias (Brodeur et al., 2024). These kinds of pressures may lead students and researchers to participating in QRPs, or in more extreme cases, even to falsifying data, or participating in other kinds of misconduct to produce favorable results (see Balafoutas et al., 2025).

A particular problem in psychology is the overreliance on the reporting of *p*-values, especially those just under 0.05 (Head et al., 2015; Masicampo & Lalande, 2012; Stevens, 2017). The overreliance may also exclude reporting the effect sizes, which is a calculated measure of the strength of a relationship between two variables within a population. Instead of reporting the effect size and describing the results with more precision, a focus is placed upon whether the data is statistically significant rather than exploring the effect. Overreliance on *p*-values indicates an incomplete description of the data. The focus on *p*-values may also lead to *p*-hacking, which is any action that intentionally manipulates *p*-values to correspond to a desired outcome. A good example of this is including more participants than originally planned. If a study currently produces approaching significant results (e.g., $p = .06$), researchers could continue recruiting participants until the desired *p*-value is reached. Using the study below, *p*-hacking would look like pretesting one language group and selecting which language gave us our desired *p*-values before recruiting a new set of participants. This type of reporting inaccurately describes the data, which may lead to different conclusions by the original researchers and those reviewing it.

Selective reporting—underreporting results, conditions, or outcomes, because they do not confirm hypotheses—is also shown to have impacted research behavior. This focuses on personal biases may influence which results are reported and which are omitted; a general dislike of reporting null findings remains in the current research climate, leading to selective reporting on statistically significant results. Selective reporting results in potentially biased data in the literature, furthering a higher risk of Type I errors in published articles (Franco et al., 2016).

1.2.2 Public trust

Trust in scientific endeavors usually relies on *epistemic trust*, trust relating to a reliable source of information (Wilholt, 2013). This trust also extends to colleagues when working on collaborative projects, by the nature of relying on others to competently complete their work and honestly report findings: *"There is no one person providing comprehensive oversight by rechecking all calculations, replicating all parts of the experiment, and so on; the background of trusting professional relationships rationalizes this practice"* (Goldenberg, 2023, pp. 369–70). Replication cannot be completed on every aspect of every study ad infinitum; we must consider practicality. Studies include some level of trust in order to practically complete new projects; replication of every study ever is an impracticality (if not an impossibility). On these grounds alone, one could argue that replication should be a health check-up for academic disciplines: *"While [replications] do not address underlying health issues (if any exist), they provide valuable insights into the state of replicability within a field"* (Balafoutas et al., 2025, p. 5). Health check-ups are periodic measurements of the state of an entity without an intense, descriptive evaluation. In this instance, a health check-up is just one study in one field. Mega-replication studies, where someone completes multiple replication studies across a field provide a health report on the field. If findings indicate a need for action, then some intervention must occur as ill findings will impact public trust in the scientific process.

This trust, however, extends to public perception of academic integrity within science, and the highest form of scientific misconduct is publishing falsified data. One example is Diederik Stapel, a social psychologist at Tilburg University in the Netherlands, who in 2011 was found to have falsified data for his graduate students. After those students generated stimuli, questionnaires, and other materials, Stapel would bring the experiment to an undisclosed location and collect the data, which he later admitted was falsified (Derksen, 2021). Deliberate fraud and other more public examples of non-replicability, such as the infamous "power pose" study—which posited that standing in a power pose for 1 minute increases testosterone, decreases cortisol, and increased feelings of power (Carney et al., 2010)—and its failed replication (Ranehill et al., 2015),

have been cited as some of the symbols of what's wrong in psychology (Derksen & Field, 2022). Fraud is the highest form of scientific misconduct, which both scientific communities and the public eye can easily understand and condemn, as conduct like this and others contributes to the erosion of public trust in scientific institutions.

Other kinds of scientific conduct and behavior can be scrutinized by the public when scientific conversation is widely available on social media. After some public displays of scientific discourse, a "tone debate" has emerged about the way scientists communicate with one another, particularly in response to the replication crisis (Derksen & Field, 2022, p. 172). When Simone Schnall's 2008 study was not replicable using much larger sample sizes (Schnall et al., 2008; Johnson et al., 2014), Schnall believed the paper should not have been published based on a flaw in the results—a ceiling effect. Schnall was in conversation with the replication authors; however, once objections regarding the statistical analyses were raised, Schnall was dismissed and the paper was published anyway. Schnall contends that this replication was not held to the same peer-review standard, which she posits would have noticed this failure. The replicators responded saying they addressed the ceiling effect and found the same non-replicable results (Schnall, 2014a). Although the objections were noted and tested, the resulting conversation online claimed replicators use "replication bullying" and create a sense of fear; for example, Schnall's graduate students worried about "data detectives" investigating their work to find something wrong (Bohannon, 2014). This spurred an intense online debate about scientific conduct in communication through social media or blog posts (Derksen & Field, 2022; Schnall, 2014b). Fiske addressed how publishing opinions outside of peer-reviewed journals are *"ignoring ethical rules of conduct because they circumvent constructive peer review: They attack the person, not just the work; they attack publicly, without quality controls"* (Fiske, 2016). This "tone debate," which has mostly taken place on social media platforms and in non–peer reviewed spaces (Derksen & Field, 2022; Fiske, 2016), is about the approach in which scientists engage in discussion about issues in research, particularly current topics like the replication crisis. For example, posts on personal blogs, Facebook, or X (Twitter) included accusations that researchers completing replication studies were "bullies" (Bohannon, 2014), "unoriginal" (Stevens, 2017), and using tactics such as shaming and

"methodological intimidation" (Fiske, 2016). The very public display of non–peer reviewed opinions resulted in action for civil discourse on any platform, discussed in more detail in the next section. An important objection to the "tone debate", however, Derksen & Field suggest that even though I began in response to the replication crisis, focusing on the interpretation and discourse of facts detracts from the heart of the crisis: the facts themselves being called into doubt (Derksen & Field, 2022, p. 179).

All of the above leads to loss of validity in published research through the public eye, ultimately diminishing epistemic trust in academic institutions and leading to further disenfranchisement in the scientific process (Balafoutas et al., 2025; Ioannidis, 2005; Wagenmakers et al., 2012). Consider the words of Daniel Sarewitz speaking about systematic errors in research: *"Nothing will corrode public trust more than a creeping awareness that scientists are unable to live up to the standards that they have set for themselves"* (Sarewitz, 2012). This distrust is only confounded as more public and non-replicable studies emerge. Losing the trust of the public poses a stark risk to our shared understanding of reality, *"leaving societies vulnerable to misinformation and unprepared to tackle urgent global challenges,"* of which the Covid-19 pandemic made acutely aware (Balafoutas et al., 2025, p. 2; West & Bergstrom, 2021). This is a call to action, to which many have started to address.

1.3 Current efforts & next steps

Many papers have offered suggestions to alleviate the replication crisis, some even suggesting that solutions cannot encompass every field and must be discipline-specific (Peterson, 2021). Some solutions, however, may be addressed to the scientific process, such as the practice of open science (e.g., preregistration), registered reports, attention to methodology and statistical analyses, publication, and student replication attempts.

Open science is the practice of transparency in research, which supplies secure accessibility to preregistrations, research plans, workflows, data, analyses, and other information pertinent to a study. It aims to diminish the impacts of investigator bias (Miguel, 2021) and hindsight bias (Wagenmakers et al., 2012), while serving as a rigorous way for researchers to avoid *p*-hacking (Stevens, 2017) and other statistical

manipulations. By securely hosting and storing these procedural materials on a digital platform (e.g., Open Science Framework [OSF]: https://osf.io/), researchers allow reviewers or replicators to follow the appropriate steps as well as being held accountable for their preregistration. A preregistration is committing aspects of a study (defining research questions, variables, experimental methods, and analysis plans) to a document before data is analyzed. This creates a roadmap for everyone organizing the study and allows for further accountability during the analysis and reporting. It is also recommended to explicitly state whether the analyses are confirmatory or exploratory (Wagenmakers et al., 2012); the purpose is to create a clear line between prediction and post-analysis (postdiction) discoveries, as to not further corrupt the data or conclusions (Nosek et al., 2018). Practices of open science still do not completely deter researchers from participating in questionable research practices; rather it makes deliberate fraudulent behavior more difficult (Balafoutas et al., 2025). Nosek et al. reflect on the importance of transparency in science: *"Openness is not needed because we are untrustworthy; it is needed because we are human"* (Nosek et al., 2012, p. 626). Preregistrations help contribute to increased transparency in research, controlling for an innate human factor: the potential for error.

Some posit that an advanced form of preregistrations are registered reports (RR) (Arpinon & Espinosa, 2023). In the first part of RR, authors submit similar information provided on a preregistration to an academic journal. Upon peer review, RR receive an in-principle acceptance, committing the journal to publish the study within a set time window, provided the study meets the standards of the journal. After IPA, researchers complete the study, analysis, and discussion and submit a second manuscript, which receives additional peer review, this time without readdressing theory, methods, or hypotheses to prevent the results from influencing recommendations (Chambers & Tzavella, 2021). RR are suggested to be an institutional solution for creating excellent research questions and methodology, and increasing transparency (Nosek et al., 2022). Since RR are selected for publication before data analyses have been completed, RR reduce publication bias and possibly the file drawer problem (Scheel et al., 2021). Objections to RR state that the practical implementation becomes worrisome. Peer reviewers are given the opportunity to comment and improve methodologies within the

RR. Suggestions from these reviewers takes time to implement and could include more expensive equipment than what the original authors may have access to. Peer review may take an excessive amount time as well, and if a study cannot start because of delay caused by RR, the research question may have become outdated in the research group or in the field, or it may have already been addressed. This institutional solution, however, remains an excellent way to improve the strength of the methodologies before data collection.

Replication studies have seemingly become more popular (Stevens, 2017) with some cognitive science journals now explicitly calling for replication studies (e.g., *Animal Behavior and Cognition*, *Perspectives on Psychological Science*). As popularity increases, additional measures should be taken to navigate original studies to their replications. Balafoutas et al., suggest a digital banner system providing links to direct or conceptual replications, fostering further transparency (2025). They also suggest that an original study's academic publisher should commit to publishing the first replication as a formal commentary, regardless of the results (Balafoutas et al., 2025). This not only combats the practice of favoring positive results, but it also furthers discourse on the research questions addressed.

As more replication studies are completed, another consideration is the complicated topic of paper retraction. Firstly, it is shown that nonreplicable studies are cited more than replicable ones (Serra-Garcia & Gneezy, 2021). If this trend is observed at large, this endangers the strength of conclusions across studies relying on the original conclusion. The biggest danger of this is when adjacent studies utilize the nonreplicable conclusions for a related study. For example, if an inclusionary or exclusionary criterium for participants is based on a paper later shown to be nonreplicable, then the acquired results are not as robust or generalizable as they should have been. To avoid this, the digital banner system suggested would help guide future researchers in an easy manner to related and important work. This is one method on how improved communication will better serve science.

Another suggestion improvement is in response to the "tone debate." The debate inspired Daniel Kahneman's proposal of "a new etiquette for replication" (2014; Schnall,

2014b). Kahneman's four rules included an original author collaborating with replicators (those conducting the replication study) with established guidelines: 1) the replicator sends detailed descriptions of the planned procedure to the author, 2) the author responds in a defined, limited period, 3) the replicator provides reasons for excluding recommendations provided by author, and 4) the entire correspondence is on the record for others to review (Kahneman, 2014). This is designed as a solution for the improper behavior shown through the "tone debate" but also approaches replication as a conversation between all researchers. This also allows original authors to participate in experiment design and analysis planning during the replication process, providing valuable insight into the original paper's design.

One suggested improvement to methodological considerations is to include multiple hypotheses before the data has been collected. This can provide researchers with stronger inferences and more ways of observing the data once it has been collected (Stevens, 2017). One way to improve the reporting of results is to include effect sizes. Making the move from if something is or is not statistically significant, the results can show the data's breadth of generalizability (Stevens, 2017). Providing this information creates a better picture of the phenomenon observed and helps to show the strength of results.

If we can assume that publication bias favoring statistically significant results has nearly reached a solution, some statistical analyses solutions may have the opportunity to address the file drawer problem by way of controlling for Type I errors. By analyzing several published studies' data together, meta-analyses have been shown to reduce Type I errors (Dalton et al., 2012). Some have also suggested that simply adjusting the critical value in statistical significance testing from the t-ratio 1.96 to 3.02 (corresponding to change in alpha levels from 5 % to 1 %) will reduce Type I errors as well (McCrary et al., 2016). These types of statistical considerations strengthen conclusions and provide better insight into the data collected.

Although it is important to note that the implemented policies combating the replication crisis are yet to be fully evaluated (Field et al., 2024), practices like preregistrations, RR, attention to analyses, and the suggestions herein seem to be promising trajectories for

higher accountability and transparency in research. In good news though, it is suggested that the replication landscape in some fields seems to be improving (Artner et al., 2021).

One last suggestion, a solution I aim to demonstrate in this thesis, is the use of student projects as a form of replication. Perrault suggests that PhD candidates should replicate published studies to improve the state of replicability in their corresponding fields (2023). Not only is student replication an excellent tool to learn and practice research skills, but it can also acquaint students more deeply with between-study discussions that occur with research literature. This learning tool may be supported by what Perrault suggests as a repository of ready-made studies that students can digest, learn, and produce valuable insight into research (2023). Although Perrault refers to a higher academic level than the purpose of this master's thesis, I aim to create a case study on precisely this suggestion.

2 Part II: Replication in practice: A case study

To explore replication in practice, the next section will showcase replication attempts from graduate level work. Experiment 1 attempts a Finnish-language conceptual replication from a French-language study (Benjamin et al., 2023), where both examine the impact of word length on statistical language learning tasks. Experiment 2 attempts a direct replication of Experiment 1, where improvement of methodology, particularly the generated stimuli, is the main focus.

2.1 An introduction to statistical language learning

A challenge faced in both infant first language learning as well as adult second language learning is the tracking of words within continuous speech, as natural speech has no pauses to indicate word boundaries. Words are also rarely presented in isolation, and thus the segmentation, and therefore also learning must take a different approach, using subtle acoustic markers such as pitch change, sound lengthening, and slowing speech rate (Benjamin et al., 2023). Adults use lexical knowledge when retrieving native-language words (Mattys et al., 2005), but in an unknown language, segmenting the words cannot rely upon previous knowledge. Study designs can mimic language learning by using novel pseudowords to test the language learning process using statistical language learning (SLL).

SLL is the process of using a theoretical set of probabilities to track linguistic regularities in an environment. Learning is measured by presenting to participants an audio (or visual) stream of linguistic stimuli (naturally occurring or novel) and testing through different tasks (explored in more detail below). Although these tasks may be completed in the visual domain, auditory examples are used throughout to accompany the experiments below.

One way SLL is measured is through transitional probabilities (TP), which are the likelihoods that a sound follows another sound. TPs are computed by calculating the frequency of X then Y divided by the overall frequency of X. Using a linguistic example, in the expression “Such a cute baby! Is it your baby?”, if given /beɪ/, what is the likelihood it is followed by /bi/, creating the word /ˈbeɪ.bi/ |baby|?

$$\frac{\text{Frequency of (XY)}}{\text{Frequency of (X)}} \; or \; \frac{\text{Frequency of /beɪbi/}}{\text{Frequency of /beɪ/}} \text{ or } P(/bi/ \mid /beɪ/) \quad (1)$$

TPs are the theoretical foundation of word segmentation, the ability to track word boundaries during SLL tasks (Saffran, Newport, et al., 1996). Word segmentation tasks test whether participants successfully track the transitions between words, and resultingly track the word boundaries. Using the example from above, if /ˈbeɪ.bi/ was included in a stream with three other words (e.g. /ˈpiː.sə/, /ˈtɑː.koʊ/, and /ˈkɑː.fi/) in a random order, the TP for the two syllables in /beɪbi/ would remain 1, whereas the boundary between words (e.g., /fi/+/beɪ/; /bi/+/pi/) would result in a drop to 0.33.

To test SLL, Saffran, Aslin, and Newport (1996) tested 8-month-old infants using four trisyllabic pseudowords in a random order within a 3-minute continuous audio stream with flat intonation. They revealed that the infants were able to distinguish between stimuli in two conditions: Words (TP for $A_1B_1C_1$ within the stream was 1 for each syllable pair) and PartWords (TP for $B_1C_1A_x$ was 1 and 0.33). By successfully segmenting the words, this showed successful tracking of TPs even in infancy. This was successfully replicated in adults by the same research group (Saffran, Newport, et al., 1996).

The results have been replicated (Black & Bergmann, 2017; Isbilen & Christiansen, 2022), shown in non-linguistic stimuli (Saffran et al., 1999; Schön et al., 2008) and the visual domain (Fiser & Aslin, 2002), and more have tested the limits of the effect. Some prosodic elements, which are speech units such as stress, rhythm, loudness, or intonation, have been used to test the extent of SLL through prosodic cues, which provide anchors to incoming stimuli, for example as word boundaries (see e.g., Kuuluvainen et al., 2025).

The typical measurement used is a two-alternative forced choice (2AFC) task, where participants must choose between two options (Word vs. PartWord) and select which one is more familiar. Using the differences in TPs, Words would become more familiar over PartWords based on how many times they appeared in the exposure stream. These scores are collected to create an accuracy score on how well participants segmented the Words within the exposure stream. In some studies, each 2AFC trial is accompanied by

a confidence rating (typically: guess, familiar, or know). This rating gathers information regarding the participants explicit or implicit knowledge of their learning. If participants correctly select the Word in the 2AFC trial and they select “know” in the confidence rating, we can conclude they have explicitly learned the Word. If they select “guess” and still correctly select the Word more times than would be expected via mere guessing, we can conclude they have implicitly learned. Once there is an increase in repetitions of stimuli presented to participants, this paradigm is shown to have decreases on correct identifications as the participants become familiar with PartWords (Soares et al., 2023).

An alternative form of testing is using a Likert scale to test familiarity of isolated words. Instead of pitting two words against each other, this test presents one stimulus at a time to retrieve how familiar the participant is. The upside is presenting each stimulus once and thus reducing learning effects. One downside is that we no longer separate accuracy and confidence, and this effectively removes direct measurement on explicit vs implicit learning. One could argue that explicit learning would resemble higher scores of familiarity, but those scores could still represent implicit learning; it becomes difficult to conclude one or the other. Familiarity ratings, however, establish scores for participants that build relationships between the two groups of stimuli.

2.1.1 Prosody and trisyllabic formations

Previous studies have shown the presence of a pause significantly improves SLL performance (Buiatti et al., 2009; Peña et al., 2002). Even though it is shown to be effective, participants do not explicitly hear the presence of the pauses, and thus no explicit acoustic cue is given (Benjamin et al., 2023; Peña et al., 2002). Not hearing the pauses explicitly yet still performing better on tasks with its presence indicates that pauses implicitly aid learning during SLL tasks when using trisyllabic word formations.

During testing, the last two sounds from the first word followed by one sound from another word (e.g., $B_1C_1A_2$) are typically used as PartWords. These are tested against the Words ($A_1B_1C_1$) to measure whether participants learned. This formation highlights the word boundaries as the placement of learning; if participants accurately segment Words

from PartWords, it will be because they have accurately tracked TPs and learned the lower TP at the word boundary.

2.1.2 Quadrisyllabic formations and Benjamin et al., 2023

Trisyllabic studies have been the foundation of SLL. From the first paradigm (Saffran, Aslin, et al., 1996) to cross-linguistic conceptual replications (Ordin et al., 2017) most SLL studies have used trisyllabic words(Benjamin et al., 2023; Isbilen & Christiansen, 2022). To test the limits of word length, Benjamin and colleagues, completed a study in infants and adults using quadrisyllabic words and 25 millisecond pauses (2023). The study created the first audio stream concatenating four artificial language quadrisyllabic words, pseudo-randomly concatenated without any pauses, and the second audio stream with pauses placed between words. The infant study was aimed at neural entrainment, which is the synchronization of neuronal activity to the rhythms of incoming stimuli. Results indicate that neural entrainment in infants during quadrisyllabic-word SLL tasks is only aided by a 25 millisecond pause. In the behavioral adult study, learning was also aided by the presence of a pause: Results indicated that quadrisyllabic word formations were only successfully segmented with the presence of a pause within the audio stream, as indicated by higher familiarity rating scores in Words than PartWords in the paused condition. The Words consisted of four syllables ($A_1+B_1+C_1+D_1$), which were used in the exposure streams to test participant's learning. PartWords consisted of the boundaries between Words in the exposure stream ($C_1+D_1+A_2+B_2$) which were heard during the audio stream, albeit with a lower frequency. ShuffleWords are word formations that never occurred during the audio stream; they are composed by switching the places of the two interior syllables ($A_1+C_1+B_1+D_1$). The difference between Words and ShuffleWords measured whether or not participants tracked TPs, since ShuffleWord TPs never existed in the exposure stream. If familiarity ratings were higher for Words than ShuffleWords, it showed they tracked TPs. Successful segmentation of quadrisyllabic Words must have shown statistically different from both PartWords and ShuffleWords. In the continuous condition in Benjamin et al., Words and PartWords did not differ statistically significantly even though Words and ShuffleWords did. This means that although

participants tracked TPs (Words–ShuffleWords), they did not successfully segment the quadrisyllabic word forms (Words–PartWords).

In Benjamin et al. (2023), each syllable was 250 milliseconds with flat intonation and no coarticulation between the sounds. A 3.3-minute audio stream was created using an artificial monotonous voice and randomly concatenating words with the only restriction that the same word could not be presented twice in a row. An additional stream (3.4 minutes) was created by adding 25 millisecond pauses between each word. All streams were ramped up and down for the first and last 5 seconds, to avoid perceptual anchors. To control for phonological similarity, Words and PartWords were reversed for half of the subjects.

After the exposure streams, participants rated familiarity of words (from “Completely unfamiliar” to “Completely familiar” on a six-step Likert scale). Exposure and test phases were completed twice. Six conditions were used to avoid any bias based on the length of words (3 bi-syllabic groups as foils, 3 quadrisyllabic groups). Participants were placed in one of two groups, controlling for order to prosodic condition presentation.

2.2 Research topic and purpose

The purpose of this thesis is to serve as a case study on replication using statistical language learning. Since it is suggested that replications completed by students can produce valuable information (Perrault, 2023), this thesis will explore replication and the replication crisis using the experiences across the two kinds of replication: conceptual and direct. Experiment 1 is a Finnish-stimuli conceptual replication of Benjamin et al. (2023). Experiment 2 is a (mostly) direct replication of Experiment 1, where focus was on improving methodology, strengthening controls, and attempting replication. In terms of SLL, these studies explore the potential impact of word length expectations on SLL tasks and if native language word length impacts those expectations in pseudowords. Experiment 2 also aims to strengthen control measures by using Titone et al.’s (2024) suggestions of confounding factors in SLL stimuli creation.

To investigate SLL in Finnish, we must first compare French and Finnish word length using graphemes and phonemes. Graphemes are the smallest written linguistic unit

(e.g., one letter in the alphabet) and phonemes are the smallest acoustic linguistic unit (e.g., /k/, /a/, /t/). Finnish has a shallow orthography, wherein each phoneme is mapped to only one grapheme (with the exception of |nk| and |ng|). French, however, has a deep orthography, wherein one grapheme (e.g., |e|) may represent multiple phonemes (e.g., /ɛ/ in |est|; /ə/ in |repli|; /e/ in |mangez|) and one phoneme (e.g., /k/) may represent multiple graphemes (e.g., |c| in |sacré|; |cc| in |accord|; |q| in |cinq|) (Besner & Smith, 1992). Therefore grapheme-defined comparison of the two languages proves difficult if not entirely unfair (see Kelih, 2012). Nevertheless, averages of grapheme-defined word-length for Finnish (7.66) and French (4.97) show a distinct difference (Kalimeri et al., 2015). Finnish also has higher uniformity in word length throughout sentence duration, indicated by word-length entropy (see Kalimeri et al., 2015) and have a relative trend for using longer words (Berg et al., 2022). Taken together, these studies suggest that on average, Finnish has longer words than French.

These led to our two hypotheses:

H1: If average word length in native speakers indicates word length expectations, and these expectations are tied to segmentation of unfamiliar speech streams based on transitional probabilities alone (statistical language learning), Finnish speakers will show greater post-exposure familiarity to quadrisyllabic word-forms embedded in a familiarization stream than to part-words from the same stream, which span across word boundaries, even when the stream is continuous. This is in contrast to the results of Benjamin et al. (2023) with speakers of French (which has much lower average word length and thus greater lower expectations for word length), who could not differentiate familiarized word-forms from part-words when the stream was continuous.

H2: If subliminal pauses aid word segmentation in Finnish speakers, similar to previously shown in French speakers (Benjamin et al., 2023), Finnish speakers would show greater post-exposure familiarity to quadrisyllabic word-forms embedded in a familiarization stream with pauses between words than to quadrisyllabic word-forms embedded in a familiarization stream with no pauses between words.

2.3 Experiment 1

Experiment 1 was conducted as part of the Experimental Lab Course (LDA-EXP315) between October 2023 and May 2024 in collaboration with Sami Huttunen and Freja Kivikettu under the advisement of Soila Kuuluvainen. The purpose of this course is to orient students to experimental design, empirical research, and exposure to the scientific process. Due to time constraints, however, we could not dedicate as much time as we wanted to all aspects (e.g., stimuli creation). Experiment 1 represents using replication as a teaching opportunity with a group of participants.

2.3.1 Methods

2.3.1.1 Participants

40 native Finnish speakers completed the study. Participants were recruited through an email list of candidates who participated in a similar experiment in 2022 and consented to be contacted again. All data was collected on April 16, 2024. Participants' data was excluded from the final report based on a predetermined set of criteria: 1) response indicating a hearing loss/condition, 2) failure of the syllable test, indicating technological problems, 3) exceptionally poor performance on the digit span test by receiving the lowest score possible of 3, indicating inattention to stimuli, or 4) reporting to have insomnia or other sleeping disorder, indicating potential atypical cognitive function. One participant was excluded for reporting a hearing loss diagnosis, and another participant was excluded for inattention indicated by exceptionally poor performance on the digit span test (confirmed by single-number responses reported). Data from the remaining 38 participants (aged 18–39, mean age 25, 17 self-identified men, 1 other) were included in analysis. This study had permission by the Ethical Review Board in the Humanities and Social and Behavioral Sciences at the University of Helsinki.

2.3.1.2 Stimuli

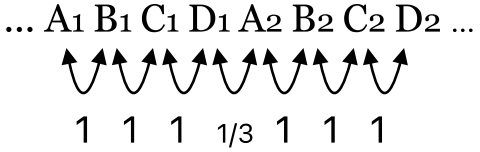

Figure 1 Transitional Probabilities
Within each quadrisyllabic word, the transitional probability (TP) between each sound was 1. At the word boundaries, the TP was 1/3.

A significant feature of the Finnish language is its usage of vowel harmony—a strict phonetic structure of word formation where front vowels (/y/, /æ/, /ø/) and back vowels (/u/, /o/, /ɑ/) cannot occur in the same word, but neutral vowels /i/ and /e/ can be used in either formation (Suomi et al., 1997). This same structure is also suggested to provide useful segmentation cues for Finnish speakers (Suomi et al., 1997), resulting in our study controlling for this feature. To accommodate for Finnish vowel harmony, two artificial languages were used in each experiment: a back artificial language, which consisted of back and neutral vowels, and a front artificial language, which consisted of front and neutral vowels.

Following the design established by Benjamin et al., we created six word groups (3 quadrisyllabic, 3 bi-syllabic) each consisting of four novel word-formations that mimic Finnish phonology. The Words consisted of four syllables (A_1+B_1+C_1+D_1), which were used in the exposure streams to test participant's learning. PartWords consisted of the boundaries between Words in the exposure stream (C_1+D_1+A_2+B_2) which were heard during the audio stream, albeit with a lower frequency. ShuffleWords are word formations that never occurred during the audio stream; they are composed by switching the places of the two interior syllables (A_1+C_1+B_1+D_1). The bi-syllabic foils were used only for testing as to not bias word length expectations for the upcoming streams. These were composed of syllables from the quadrisyllabic options: 4 selected from Words (C_1+D_1), 4 selected from PartWords (D_1+A_2), and 4 selected from ShuffleWords (C1+B1). When creating these new stimuli, we needed to consider the

formations of Words, PartWords, and ShuffleWords, and we used the following criteria for both experiments. 1) The Words contain four syllables and each syllable has preestablished criteria: a) the syllable must exist within Finnish linguistic constraints (e.g., /q/ is not a native Finnish letter and therefore not considered), b) the syllable must not contain comprehensible linguistic information such as free morphemes (ja, jo, mä, sä, me, se, te, he, ne), bound morphemes (-ni, -si, -ko, -kö, -pa, -pä, -jä, -na, -nä, -ta, -tä, -da, -dä), or exclamations or abbreviated words common in spoken Finnish (ha, ho, hi, hä, hö, ku). 2) Within each Word, PartWord, and ShuffleWord, the combinations of syllables must not create comprehensible morphemes (e.g., exclude "väli," which means "gap" or "distance").

In Experiment 1, stimuli were created from querying a large Finnish-language corpus using quadrisyllabic words and a series of trial-and-error tests to select the appropriate candidates. See Table 1 for the chosen stimuli.

2.3.1.3 Pretest

Before Experiment 1, Pretest 1 was completed to discover how similar the novel pseudowords were to Finnish words using a continuous scale from 1 to 6.[1] 7 participants (aged 22–50, mean age 37, 3 self-reported men) were recruited opportunistically through friends, family, and social media. Participants performed the digital experiment hosted on Pavlovia (https://pavlovia.org). Unfortunately, data was not collected for the ShuffleWord *risojenu*. The results of Pretest 1 were used to identify potential outliers before proceeding to Experiment 1 and to select the best PartWords for post-exposure testing. PartWords were selected by their similarity in scores to their corresponding Words and ShuffleWords. We also included the constraint that all syllables present in the Words must be present in the PartWords only once (i.e., *syreryvi* and *lyriryvi* were mutually exclusive alternatives).

[1]"Valitse asteikolla, miten suomen kielen sanalta sana kuullostaa [sic]."

Table 1 Average Pretest 1 Rating
These results were used to select the PartWords based on similarity to both Words and ShuffleWords. The selected PartWords are presented below.

Word ($A_1B_1C_1D_1$)	*Mean (Standard Deviation)*	PartWord ($C_1D_1A_2B_2$)	*Mean (Standard Deviation)*	ShuffleWord ($A_1C_1B_1D_1$)	*Mean (Standard Deviation)*
lävesyre	2.88 (0.617)	*syreryvi*	3.51 (1.121)	*läsyvere*	4.08 (0.473)
ryviröle	3.34 (0.964)	*rölevöli*	3.25 (0.816)	*ryrövile*	3.44 (1.139)
tikelyri	3.72 (1.121)	*lyriläve*	3.15 (0.778)	*tilykeri*	4.08 (0.817)
völiväty	3.71 (1.183)	*vätytike*	3.50 (0.606)	*vövälity*	3.48 (1.248)
TOTAL	**3.41 (0.760)**	**TOTAL**	**3.35 (0.520)**	**TOTAL**	**3.77 (0.563)**
rijesonu	3.44 (1.226)	*sonuvele*	3.81 (1.087)	*risojenu*	n/a
salutira	3.90 (1.310)	*tirarije*	2.71 (1.003)	*satilura*	3.87 (1.324)
velemiro	3.32 (1.173)	*mirovoli*	3.28 (0.973)	*vemilero*	3.01 (1.120)
volirelo	2.73 (0.683)	*relosalu*	3.58 (1.176)	*vorelilo*	2.78 (0.621)
TOTAL	**3.34 (0.840)**	**TOTAL**	**3.35 (0.845)**	**TOTAL**	**3.22 (0.705)**

2.3.1.4 Procedure

The Gorilla Experiment Builder was used to create and host this experiment (www.gorilla.sc; Tomczak et al., 2023).

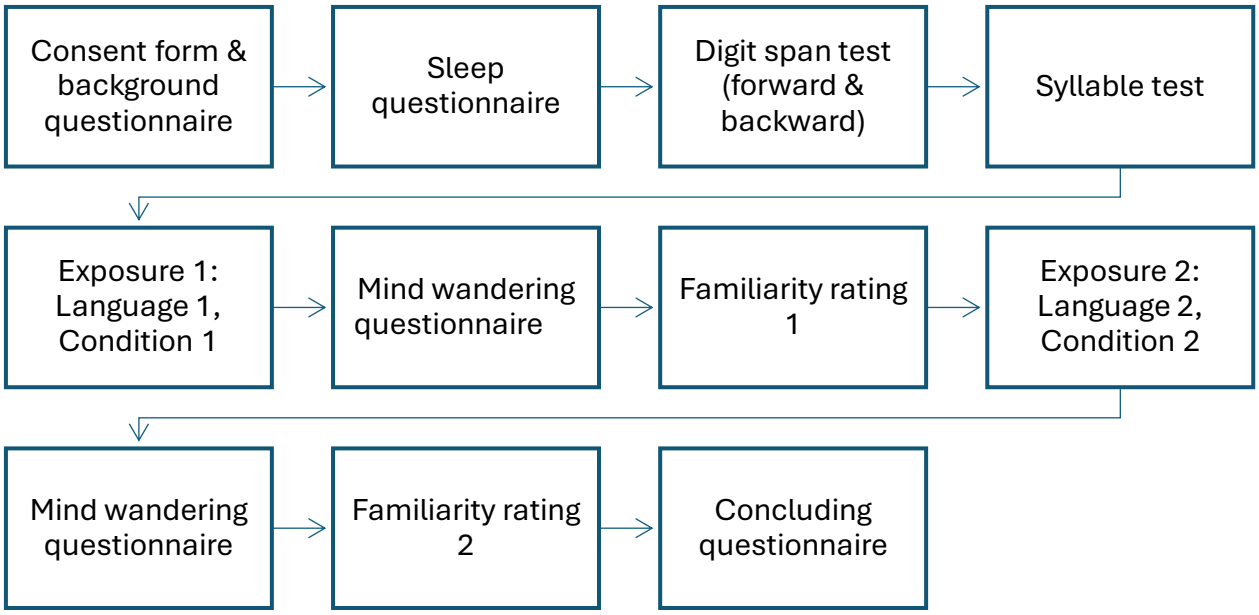

Figure 2 Procedure
Procedure for experiments 1 and 2.

In Experiment 1, participants began with a consent form, followed by a background questionnaire (age, gender, native language, number of fluent second languages, dominant hand, hearing issues, and learning disabilities). The next task to complete was a sleep questionnaire, selected questions from the Pittsburgh Sleep Quality Index (Mollayeva et al., 2016), to collect information on participant's quality of sleep during the previous night. Following that was a visual digit span test, which individually presents numbers in a series for participants to then report the digits in the order they appeared. This was then repeated for participants to report the digits in reverse order. This test was used only to verify that participants were attending to this unsupervised, digital experiment and to familiarize them with the platform, and thus the digit span data was not analyzed in relationship to working memory. Although the digit span test traditionally utilizes verbal presentation of the numbers (Bahri et al., 2024; Jones & Macken, 2015; Schroeder et al., 2012), changing the presentation from verbal to visual is suggested to be a fair assessment of working memory (Olsthoorn et al., 2014). Participants were then given a syllable test, wherein participants listened to an audio

clip of syllables and were asked to select the sounds from a list of syllables (e.g., |mi|, |ji|, |vö|, etc.). This task verified the participant's audio equipment was properly working.

Then, participants were randomized into one of four groups. Each group was given two exposure streams: the first was either the front or back language with either the paused or the continuous condition; the second exposure stream was the opposite of both features (e.g., Group 1: Stream 1 = Paused Front; Stream 2 = Continuous Back).

Participants were told in Finnish: "Next you will hear a series of syllables. It will last about three minutes. Try to listen with concentration."[2] After the three-minute exposure, participants were given a short mind wandering questionnaire (see Appendix: Mind Wandering Questionnaire) to collect information about attention and mind wandering, not analyzed here. Then participants were asked to give familiarity ratings on a Likert scale from 1 to 6: "Next, you will hear individual short words. Your task is to judge how familiar the word sounds. Be careful, you will only hear each word once."[3] Participants were given all six word groups (3 quadrisyllabic, 3 bi-syllabic) in randomized order to give familiarity ratings, totaling 24 trials. Once completed, participants listened to the second exposure stream, followed by a mind wandering questionnaire and familiarity rating. Finally, participants were given a concluding questionnaire, which asked about audio quality, gave opportunity for additional comments, and provided the completion code and instructions for compensation. On average, the whole experiment lasted 20 minutes.

2.3.1.5 Statistical analysis

The following analyses were performed: First, we tested for the effects of order (paused/continuous first) with a repeated measures ANOVA (rANOVA) of 2 (Condition: pause vs. continuous) x 3 (Stimulus Type: Word, PartWord, ShuffleWord) x 2 (Order: paused first vs. continuous first). Then, the effect of language was tested with an

[2] "Seuraavaksi kuulet tavusarjan. Se kestää noin kolme minuuttia. Koeta kuunnella keskittyneesti."
[3] "Seuraavaksi kuulet yksittäisiä lyhyitä sanoja. Tehtäväsi on arvioida, kuinka tutulta sana kuulostaa. Ole tarkkana, kuulet jokaisen sanan vain kerran."

rANOVA of 2 (Condition) x 3 (Stimulus Type) x 2 (Paused Language: paused back language vs. paused front language) was completed to examine the effects of Condition, Stimulus Type, and Paused Language on familiarity ratings. Finally, an rANOVA of 2 (Condition) x 3 (Stimulus Type) was performed to test the effect of pause on familiarity ratings. Significant interactions were investigated further with post hoc tests using a Bonferroni correction. All analyses were performed with IBM SPSS Statistics (Version 29.0.2.0).

2.3.2 Results

The rANOVA of Condition x Stimulus Type x Order revealed a significant main effect of Stimulus Type $F(2,72) = 36.0$, $p < .001$, $\eta_p^2 = .500$. According to post hoc tests, Words were rated as most familiar (M = 4.138), PartWords second (M = 3.537), and ShuffleWords third (M = 2.971), with all three types differing from each other statistically significantly ($p < .001$). There were no other significant main effects or interactions. Therefore, the order in which the conditions were presented did not affect the results.

The rANOVA of Condition x Stimulus Type x Paused Language revealed again a significant main effect of Stimulus Type $F(2, 72) = 37.60$, $p < .001$, $\eta_p^2 = .511$, which indicates again that all three types of stimuli differ from each other statistically significantly ($p < .001$), with Words most familiar (M = 4.155), PartWords second most familiar (M = 3.549), and ShuffleWords third (M = 2.980). There were no significant main effects for Condition ($p = .94$) or Paused Language ($p = .76$). However, there was a significant three-way interaction of Condition x Stimulus Type x Pause Language $F(2, 72) = 4.46$, $p = .015$, $\eta_p^2 = .110$. There were no other significant main effects or interactions.

Further examination of the three-way Condition x Stimulus Type x Pause Language interaction with Bonferroni-adjusted pairwise comparisons revealed significant differences among stimuli within each language and condition (Figure 3). While the difference for paused Front Words (M = 4.329) and paused Front PartWords (M = 3.763) merely approached significance ($p = .065$) with a mean difference of 0.566, 95%

CI [-0.026, 1.157], participants showed significantly (p = .004) higher familiarity for paused Back Words (M = 3.947) than paused Back PartWords (M = 3.132), with a mean difference of 0.816, 95% CI [0.224, 1.407]. In the continuous conditions, participants showed significantly (p = .004) higher familiarity for continuous Back Words (M = 4.263) than continuous Back PartWords (M = 3.474), with a mean difference of 0.789, 95% CI [0.216, 1.363], whereas there was no statistical significance (p = .842) between continuous Front Words (M = 4.079) and continuous Front PartWords (M = 3.829), with a mean difference of 0.250, 95% CI [-0.323, 0.823]. The difference between Words and ShuffleWords was statistically significant in every instance for both languages and conditions (.001 < p < .039).

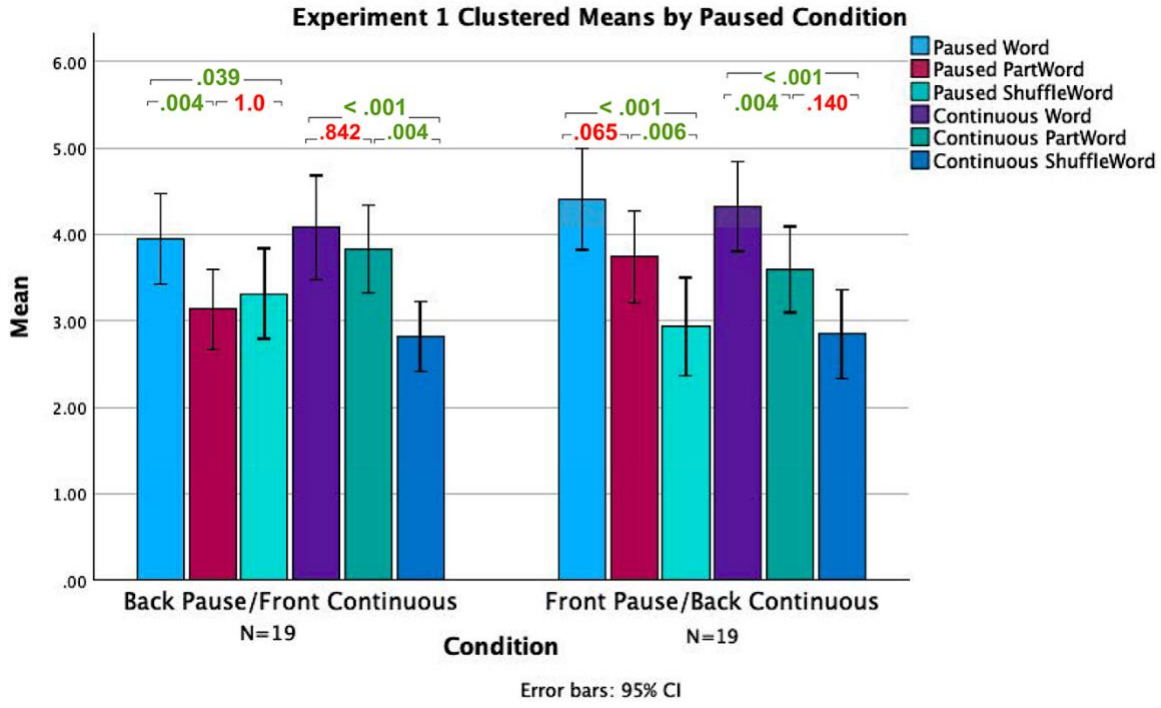

Figure 3
Results from Experiment 1 between languages and conditions. The Front language results are the interior six bars; the Back language results are the six exterior bars. The p-values are listed above the bars.

A final analysis combining the data from the two languages was conducted regardless of the statistically significant interaction of Condition x Stimulus Type x Paused Language. This decision was made to increase the comparability of the results with those of Benjamin et al. 2023 as well as increase the statistical power of the analysis by including

the full set of participants (N=38) in the analysis for the effect of pauses on the familiarity ratings. A 2 (Condition) x 3 (Stimulus Type) rANOVA revealed a statistically significant main effect of Stimulus Type ($F(2, 74) = 37.13$, $p < .001$, $\eta_p^2 = .501$), with Words rated highest, PartWords second highest and ShuffleWords lowest. Ratings for all three Stimulus types differed from each other (p<.001). Since there was no significant interaction of Condition x Stimulus Type, the results suggest that Finns can segment quadrisyllabic streams equally well with and without a pause.

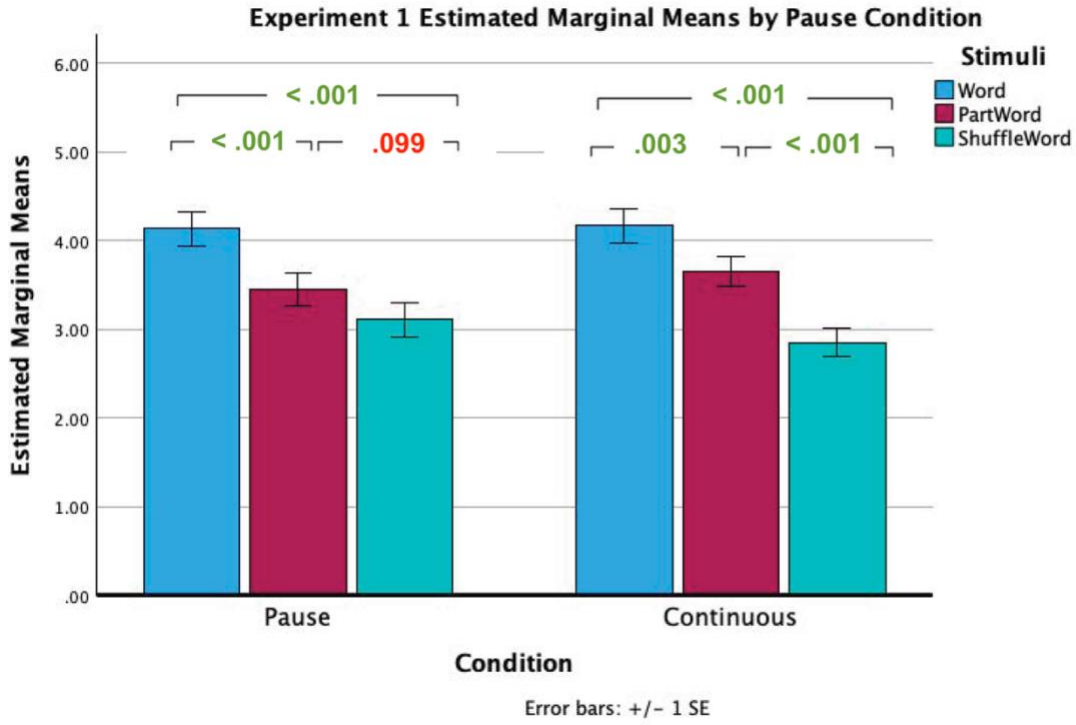

Figure 4
Results from Experiment 1 between conditions. The p-values are listed above the bars.

2.3.3 Interim discussion

Unlike what we predicted, the two language groups behaved differently during testing. Resultingly, we analyzed the language groups separately, reducing our statistical power. These results indicate a failed replication in both languages. Where Benjamin et al. found failure to segment in the continuous condition, we also found failure to segment in the continuous Front Words ($p = .842$) but a successful segmentation in the continuous Back Words ($p = .004$). Where Benjamin et al. found successful

segmentation in the paused condition ($p < .001$), we saw a failure of segmentation approaching significance in the paused Front Words ($p = .065$), but we found a successful segmentation in the paused Back Words ($p = .004$). In summary, the results suggest that when participants received the continuous condition in the Front language they could not differentiate between Words and PartWords (similarly to Benjamin et al. 2023), whereas when receiving it in the Back language they could make this distinction (supporting the word-length expectation hypothesis).

After discovering an effect in the data caused by which language was given, it necessitated categorizing scores into language-specific groups for analysis. The difference in languages may indicate a vowel-placement effect on word segmentation; however, since the groups were so small, it diminished the statistical power given to analyses and drawing such conclusions warrants more support. We completed an analysis combing all scores, which resulting in support of our hypothesis that Finnish speakers can segment quadrisyllabic word forms from continuous speech. However, as illustrated above, the differences between each language indicate that these results are not reliable.

Another limitation outside small group sizes is the incomplete attention given to stimuli creation during the early stages of the study. The stimuli creation completed was satisfactory for the original intent of this study— a course assignment requiring a focus on scientific collaboration and the scientific process. However, stronger linguistic controls should have been implemented as to not potentially influence results. One possibility was proposed by Titone et al. (2024) who suggested controls for confounding factors the stimuli may have on SLL, including native language semantics, TP appearances, phonological patterns, and acoustic variations.

Detected errors are another potential limitation of this study: An error was identified in the audio file for the syllable “vö” in the front language. In Experiment 1, the length of this sound was .251 seconds, instead of the uniform .222. This error may have resulted in an additional cue to participants, separating “vö” as a separate sound and transforming the quadrisyllabic Word (*völiväty*) into a trisyllabic Word (*liväty*). This 29-millisecond difference may have acted as an additional subliminal pause and

confounded learning of the quadrisyllabic sequence in both conditions. This may have influenced how poorly participants performed on this task, which requires fixing and validating results. This error was corrected for Experiment 2.

2.4 Experiment 2

Addressing the limitations from Experiment 1, Experiment 2 was aimed to be a direct replication of Experiment 1 with improved controls for potential impacts of the stimuli. Each aspect of Titone et al.'s suggestions were addressed and applied directly to the stimuli from Experiment 1 to confirm or deny if they are viable options to consider (2024). Whereas Experiment 1 was using replication as a teaching tool in a group setting, Experiment 2 uses replication as a teaching tool to improve individual scientific practices.

2.4.1 Methods

2.4.1.1 Participants

45 native Finnish speakers completed the study. Participants were recruited through email and data was collected on March 7–23, 2025. The exclusionary criteria from Experiment 1 were intensely reviewed to confirm each usage. After confirmation, two additional exclusionary criteria were included for this second study: 1) a self-declared diagnosis of insomnia or other sleeping disorder and 2) high musical ability (more than 10 years of musical training of an instrument or singing). Insomnia exclusion was based on: a) the hyperarousal model of insomnia positing an increase in cognitive activity as typical to the disorder (Riemann et al., 2010) and thus abnormal cognitive function potentially impacting results, and b) the suggested link between working memory and sleep discrepancy (Musich et al., 2024), requiring an exclusion for sleep-related disorders. High musical ability was excluded since evidence suggests that musicians have better auditory statistical learning than non-musicians (Mandikal Vasuki et al., 2016). The threshold of 10 years of musical training has been used for several musical ability–based studies (Ruggles et al., 2014; Strait et al., 2010). 8 participants were excluded due to musicality scores, 1 was excluded for a sleep disorder diagnosis, and 1 was excluded for both musicality and a sleep disorder diagnosis. The final sample

consisted of data from 35 participants (aged 18–31, mean age 24.7, 16 self-identified men, 1 other). The study had permission by the Ethical Review Board in the Humanities and Social and Behavioral Sciences at the University of Helsinki.

2.4.1.2 Stimuli

In Experiment 2, particular attention was paid to stimuli creation. The stimuli were designed following recommendations by Titone et al. for stimuli in SLL tasks, wherein the following confounding factors were addressed: native language statistics, phonological rhythms, between-word transitional probabilities, and acoustic rhythms of audio streams (Titone et al., 2024). Their article suggests a series of codes to streamline investigation into trisyllabic word formations for these tasks. Since we were unable to successfully manipulate the code to accommodate quadrisyllabic word formations, these factors were manually reviewed in detail to control for potential influences on results.

2.4.1.2.1 Native language statistics

The first confounding factor considered is the native language statistics. This is meant to investigate the probabilities that each sound is represented in each language's corpora. In order to control for this, we queried a Finnish language corpus investigating sounds at the bi-, tri-, and quadrisyllabic word level. We used the aforementioned linguistic constraint criteria and investigated the placement of each sound within a quadrisyllabic word. By isolating each syllable and querying the corpus, we were able to create a chart of number of appearances for each syllable within quadrisyllabic words within a Finnish linguistic corpus. Placing the frequencies against each other in those positions, we were able to generate a percentage of how many times certain sounds exist in each position of a quadrisyllabic word in Finnish. These statistics are then used to verify positions within each candidate Word. Some studies aim to exclude syllables associated with high frequencies (Kiai & Melloni, 2021) and others aim to normalize syllable-level positional similarity relative to the known language (Gómez Varela et al., 2024). For our study, we used the later and aimed for positional frequencies between 20–45%. Considering the difficulty of creating a perfect solution to this constraint, we settled for the best possible solution—a constraint which should be considered.

Table 2 Positional Frequencies
Positional frequencies are calculated by first querying quadrisyllabic words from a linguistic corpus. Number of appearances for each syllable are collected for each position within the quadrisyllabic words and the distribution is recorded in percentages. (Red = ≤ 0.10 or ≥ 0.45. Yellow = 0.10 < x ≤ 0.20. Blue = 0.20 < x < 0.45.)

Front AL	1st	2nd	3rd	4th	**Experiment 1 Back AL**	1st	2nd	3rd	4th	**Experiment 2 Back AL**	1st	2nd	3rd	4th
sa	0.35	0.27	0.16	0.22	*vö*	0.03	0.60	0.31	0.06	mi	0.19	0.32	0.35	0.14
lu	0.30	0.28	0.29	0.13	*li*	0.26	0.31	0.32	0.10	ru	0.42	0.39	0.17	0.03
ti	0.12	0.32	0.36	0.20	*vä*	0.22	0.20	0.45	0.13	vo	0.33	0.41	0.19	0.06
ra	0.42	0.35	0.19	0.04	*ty*	0.23	0.27	0.42	0.08	na	0.18	0.28	0.17	0.37
vo	0.33	0.42	0.19	0.06	*ry*	0.37	0.28	0.27	0.08	re	0.38	0.28	0.30	0.04
li	0.26	0.31	0.32	0.10	*vi*	0.26	0.30	0.26	0.17	ji	0.04	0.20	0.48	0.27
re	0.39	0.28	0.30	0.04	*rö*	0.32	0.44	0.23	0.01	nu	0.26	0.25	0.42	0.07
lo	0.26	0.39	0.25	0.10	*le*	0.34	0.29	0.33	0.05	so	0.41	0.30	0.22	0.06
ve	0.40	0.39	0.18	0.04	*lä*	0.51	0.17	0.26	0.06	le	0.34	0.29	0.33	0.04
le	0.34	0.29	0.33	0.05	*ve*	0.40	0.39	0.18	0.04	ju	0.52	0.28	0.14	0.06
mi	0.18	0.32	0.35	0.14	*sy*	0.58	0.26	0.15	0.02	sa	0.35	0.27	0.16	0.22
ro	0.39	0.34	0.21	0.06	*re*	0.39	0.28	0.30	0.04	pu	0.50	0.29	0.17	0.04
ri	0.19	0.28	0.43	0.10	*ti*	0.12	0.32	0.36	0.20	vu	0.40	0.37	0.19	0.04
je	0.15	0.25	0.58	0.02	*ke*	0.36	0.45	0.14	0.05	mo	0.37	0.38	0.20	0.06
so	0.41	0.31	0.22	0.06	*ly*	0.26	0.24	0.32	0.18	ke	0.37	0.44	0.13	0.05
nu	0.26	0.25	0.42	0.07	*ri*	0.19	0.28	0.43	0.10	ti	0.13	0.32	0.35	0.20

2.4.1.2.2 Phonological rhythms

The second confounding factor, phonological rhythms, was tested in Experiment 2 by categorizing each syllable based on its phonological features. Titone et al. suggest aspects for consonants (sonorant, continuant, lateral, nasal, voiced, coronal, labial, high, back) and vowels (labial, high, low, back) (2024). Table 3 shows the pattern for phonological features for Experiment 1.

Table 3 Experiment 1 Phonological Feature Table
Phonological features from the front- and back-vowel artificial languages for Experiment 1. Potential confounding patterns are highlighted in blue.

Front AL	*vö*	*li*	*vä*	*ty*	*ry*	*vi*	*rö*	*le*	*lä*	*ve*	*sy*	*re*	*ti*	*ke*	*ly*	*ri*
Sonorant	0	**1**	0	0	**1**	0	**1**	**1**	**1**	0	0	**1**	0	0	**1**	**1**
Continuant	**1**	**1**	**1**	0	0	**1**	0	**1**	**1**	**1**	**1**	0	0	0	**1**	0
Lateral	0	**1**	0	0	0	0	0	**1**	**1**	0	0	0	0	0	**1**	0
Nasal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Voiced	**1**	**1**	**1**	0	**1**	**1**	**1**	**1**	**1**	**1**	0	**1**	0	0	**1**	0
Coronal	0	**1**	0	**1**	**1**	0	**1**	**1**	**1**	0	**1**	**1**	**1**	0	**1**	**1**
Labial	**1**	0	**1**	0	0	**1**	0	0	0	**1**	0	0	0	0	0	0
High	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Back	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0
Labial	**1**	0	0	**1**	**1**	0	**1**	0	0	0	**1**	0	0	0	**1**	0
High	0	**1**	0	**1**	**1**	**1**	0	0	0	0	**1**	0	**1**	0	**1**	**1**
Low	0	0	**1**	0	0	0	0	0	**1**	0	0	0	0	0	0	0
Back	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Back AL	*sa*	*lu*	*ti*	*ra*	*vo*	*li*	*re*	*lo*	*ve*	*le*	*mi*	*ro*	*ri*	*je*	*so*	*nu*
Sonorant	0	**1**	0	**1**	0	**1**	**1**	**1**	0	**1**	**1**	**1**	**1**	0	0	**1**
Continuant	**1**	**1**	0	0	**1**	**1**	0	**1**	**1**	**1**	0	0	0	0	**1**	0
Lateral	0	**1**	0	0	0	**1**	0	**1**	0	**1**	0	0	0	0	0	0
Nasal	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	**1**
Voiced	0	**1**	0	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	0	**1**
Coronal	**1**	**1**	**1**	**1**	0	**1**	**1**	**1**	0	**1**	0	**1**	**1**	0	**1**	**1**
Labial	0	0	0	0	**1**	0	0	0	**1**	0	**1**	0	0	0	0	0
High	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0
Back	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Labial	0	**1**	0	0	**1**	0	0	**1**	0	0	0	**1**	0	0	**1**	**1**
High	0	**1**	**1**	0	0	**1**	0	0	0	0	**1**	0	**1**	0	0	**1**
Low	**1**	0	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0
Back	**1**	**1**	0	**1**	**1**	0	0	**1**	0	0	0	**1**	0	0	**1**	**1**

This table shows potential rhythms that may inadvertently aid learning during the exposure. For instance, in the Back language in Experiment1, the first two syllables of three Words created an accidental pattern of sonorant, continuant, and lateral features. This pattern may have unintentionally indicated to participants that these were the starting sounds to Words. In those same Words, the last two syllables also contained vowels in a high-then-back formation, which also may have aided learning. These patterns illustrate the potential impact of phonological features on learning.

For Experiment 2, we generated this chart for each candidate Word and investigated potential confounding phonological patterns. As seen below, patterns still emerge, albeit small; this control is meant to highlight potential negative impacts of the stimuli, if any are present. In Experiment 2, the back artificial language has a repeating pattern within the vowels; however, the position of the pattern does not produce additional constraints as it did in Experiment 1. This verifies that no unintentional pattern is present in the stimuli.

Table 4 Experiment 2 Phonological Feature Table
Phonological features from the front- and back-vowel artificial languages for Experiment 2. Potential confounding patterns are highlighted in blue.

Front AL	*vö*	*li*	*vä*	*ty*	*ry*	*vi*	*rö*	*le*	*lä*	*ve*	*sy*	*re*	*ti*	*ke*	*ly*	*ri*
Sonorant	0	**1**	0	0	**1**	0	**1**	**1**	**1**	0	0	**1**	0	0	**1**	**1**
Continuant	**1**	**1**	**1**	0	0	**1**	0	**1**	**1**	**1**	**1**	0	0	0	**1**	0
Lateral	0	**1**	0	0	0	0	0	**1**	**1**	0	0	0	0	0	**1**	0
Nasal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Voiced	**1**	**1**	**1**	0	**1**	**1**	**1**	**1**	**1**	**1**	0	**1**	0	0	**1**	0
Coronal	0	**1**	0	**1**	**1**	0	**1**	**1**	**1**	0	**1**	**1**	**1**	0	**1**	**1**
Labial	**1**	0	**1**	0	0	**1**	0	0	0	**1**	0	0	0	0	0	0
High	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Back	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0
Labial	**1**	0	0	**1**	**1**	0	**1**	0	0	0	**1**	0	0	0	**1**	0
High	0	**1**	0	**1**	**1**	**1**	0	0	0	0	**1**	0	**1**	0	**1**	**1**
Low	0	0	**1**	0	0	0	0	0	**1**	0	0	0	0	0	0	0
Back	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Back AL	*mi*	*ru*	*vo*	*na*	*re*	*ji*	*nu*	*so*	*le*	*ju*	*sa*	*pu*	*vu*	*mo*	*ke*	*ti*
Sonorant	**1**	**1**	0	**1**	**1**	0	**1**	0	**1**	0	0	0	0	**1**	0	0
Continuant	0	0	**1**	0	0	0	0	**1**	**1**	0	**1**	0	**1**	0	0	0
Lateral	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0
Nasal	**1**	0	0	**1**	0	0	**1**	0	0	0	0	0	0	**1**	0	0
Voiced	**1**	**1**	**1**	**1**	**1**	**1**	**1**	0	**1**	**1**	0	0	**1**	**1**	0	0
Coronal	0	**1**	0	**1**	**1**	0	**1**	**1**	**1**	0	**1**	0	0	0	0	**1**
Labial	**1**	0	**1**	0	0	0	0	0	0	0	0	**1**	**1**	**1**	0	0
High	0	0	0	0	0	**1**	0	0	0	**1**	0	0	0	0	0	0
Back	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0
Labial	0	**1**	**1**	0	0	0	**1**	**1**	0	**1**	0	**1**	**1**	**1**	0	0
High	**1**	**1**	0	0	0	**1**	**1**	0	0	**1**	0	**1**	**1**	0	0	**1**
Low	0	0	0	**1**	0	0	0	0	0	0	**1**	0	0	0	0	0
Back	0	**1**	**1**	**1**	0	0	**1**	**1**	0	**1**	**1**	**1**	**1**	**1**	0	0

2.4.1.2.3 Between-word transitional probabilities

The next confounding factor considers between-word transitional probabilities within the exposure stream. This controls appearances of bi-syllabic sounds at Word boundaries as to not favor certain PartWords during testing. If left unchecked, there is a possibility that $Word_1$ would precede $Word_2$ more times than it would $Word_3$ or $Word_4$. For example, if *salutira* precedes *velemiro* more times than it does *rijesonu*, then the participant heard the part word *tiravele* more times than they did *tirarije*, resulting in more familiarity and thus impacting testing of those PartWords. To control for this phenomenon, Titone et al. suggest a pseudo-random walk, which creates a chart of bi-syllabic sounds one at a time to create the same transitional probability between each (2024).

Table 5 Experiment 1 Between-Word Transitional Probabilities
The transitional probabilities presented in the exposure streams for Experiment 1. The numbers were collected by listening to the stream, collecting the frequency of Word order pairs, and creating the probability from a pair's frequency over the total number of appearances of the 1st Word. Highlighted in blue are the between-word transitions for the selected PartWords used in the post-exposure testing.

Back Pause	2nd *salutira*	2nd *volirelo*	2nd *velemiro*	2nd *rijesonu*	Back Continuous	2nd *salutira*	2nd *volirelo*	2nd *velemiro*	2nd *rijesonu*
1st *salutira*	0	0.34	0.44	0.22	1st *salutira*	0	0.22	0.440.51	0.27
1st *volirelo*	0.36	0	0.28	0.36	1st *volirelo*	0.32	0	0.26	0.42
1st *velemiro*	0.26	0.34	0	0.40	1st *velemiro*	0.26	0.42	0	0.32
1st *rijesonu*	0.39	0.33	0.29	0	1st *rijesonu*	0.40	0.36	0.24	0
Front Pause	2nd *völiväty*	2nd *ryviröle*	2nd *lävesyre*	2nd *tikelyri*	Front Continuous	2nd *völiväty*	2nd *ryviröle*	2nd *lävesyre*	2nd *tikelyri*
1st *völiväty*	0	0.34	0.28	0.38	1st *völiväty*	0	0.30	0.34	0.36
1st *ryviröle*	0.26	0	0.42	0.32	1st *ryviröle*	0.37	0	0.29	0.35
1st *lävesyre*	0.47	0.22	0	0.31	1st *lävesyre*	0.44	0.26	0	0.30
1st *tikelyri*	0.28	0.44	0.28	0	1st *tikelyri*	0.18	0.44	0.38	0

In line with this suggestion, a JavaScript sequence using a .json file was generated using Python. In this script, it provided variables A, B, C, and D in a pseudo-random walk sequence. Using the sequence as a template, and—after considering the following confounding factor—the four selected Words' audio files were inserted in place of those variables to generate the exposure streams. This framework provided the same random sequence to both languages and to both conditions, creating uniformity across conditions and languages for further control.

Table 6 Experiment 2 Between-Word Transitional Probabilities
This example shows the controlled transitional probabilities for all four exposure streams in Experiment 2.

Experiment 2	2nd *völiväty*	2nd *ryviröle*	2nd *lävesyre*	2nd *tikelyri*
1st *völiväty*	0	0.333…	0.333…	0.333…
1st *ryviröle*	0.333…	0	0.333…	0.333…
1st *lävesyre*	0.333…	0.333…	0	0.333…
1st *tikelyri*	0.333…	0.333…	0.333…	0

2.4.1.2.4 Acoustic rhythm

The final confounding factor was acoustic rhythm, which is considered during the creation of the audio files. This is meant to control for confounding elements within the frequency envelope as to not favor certain sounds or lengths in words. This can particularly be problematic if the confounding sound is at the beginning of a word,

giving extra indication to participants that a new word has begun. This has been controlled for in the syllable creation stage, wherein sounds were selected based on their similarity with surrounding sounds, including the tone. This was completed by having one female voice speak nonsense words composed of the target syllable in the second position of a quadrisyllabic word. The vowel from the targeted syllable was repeated for every syllable, preceded by /k/ (e.g., “pu” within “kupukuku”). /k/ is a plosive, which contains empty sounds before production, which aides extracting the targeted sounds and vowels around it much easier. This quadrisyllabic word was repeated by the speaker three times and the most neutral form was selected for processing. This process was completed by other researchers.

Once syllables were selected, they were processed into .wav files using Praat (Boersma & Weenink, 2025) in this way: each syllable was compressed to .222 seconds, sampled at 48000 Hz, and concatenated with the corresponding syllables to create .888 second pseudo-words (Words). The same process generated the test stimuli groups (3 quadrisyllabic, 3 bi-syllabic) for post-exposure testing. Words were concatenated into 3-minute exposure streams (182 seconds). For the paused condition, a 25-millisecond pause was inserted at the end of each Word audio file, which resulted in a .913-second Word. The pause-Words were concatenated into a 3.1-minute exposure stream (187 seconds). Using Python, a 5-second fade-in and fade-out were added to each exposure stream to control for any perceptual anchors. The .wav files were exported using iTunes into .mp3 files to be compatible with the Gorilla platform.

2.4.1.3 Pretest 2

During Pretest 2, 13 native Finnish speakers (5 self-identified men, 2 other) aged 29–55 were recruited opportunistically via friends and family to rate presented words on a Likert scale from 1 to 6 based on how much that word sounds like a real Finnish word.[4] As noted above, the front language remained the same from Experiment 1, and the back language was updated according to the above confounding factors. Participants were presented with two groups of stimuli, one containing the back language candidates and

[4] “Valitse asteikolla, kuinka paljon sana kuulostaa suomenkieliseltä sanalta.”

the other containing the front language candidates. 6 participants received the front language group first and then the back, and 7 participants received the back language first and then the front.

Table 7 Average Pretest 2 Ratings
These results were used to select the PartWords based on similarity to both Words and ShuffleWords. The selected PartWords are presented below.

Word ($A_1B_1C_1D_1$)	*Mean (Standard Deviation)*	PartWord ($C_1D_1A_2B_2$)	*Mean (Standard Deviation)*	ShuffleWord ($A_1C_1B_1D_1$)	*Mean (Standard Deviation)*
lävesyre	3.15 (1.068)	*syreryvi*	2.77 (1.301)	*läsyvere*	2.92 (1.441)
ryviröle	2.77 (1.787)	*rölevöli*	2.77 (1.739)	*ryrövile*	2.69 (1.316)
tikelyri	3.69 (1.494)	*lyriläve*	2.92 (1.441)	*tilykeri*	4.46 (1.198)
völiväty	3.62 (1.502)	*vätytike*	2.54 (1.330)	*vövälity*	2.00 (1.225)
TOTAL	**3.31 (1.081)**	**TOTAL**	**2.75 (0.884)**	**TOTAL**	**3.01 (0.932)**
lejusapu	3.00 (1.528)	*sapumiru*	3.08 (1.382)	*lesajupu*	2.42 (1.379)
miruvona	2.92 (1.188)	*vonavumo*	2.62 (1.446)	*mivoruna*	3.08 (1.553)
rejinuso	2.46 (1.127)	*nusoleju*	2.92 (1.605)	*renujiso*	2.46 (1.391)
vumoketi	2.92 (1.498)	*ketireji*	3.31 (1.797)	*vukemoti*	2.00 (1.080)
TOTAL	**2.83 (1.096)**	**TOTAL**	**2.98 (1.116)**	**TOTAL**	**2.48 (0.927)**

The Pretest 2 results indicated which PartWords to use for post-exposure testing in Experiment 2. A 2 (Language) x 3 (Stimuli) repeated measures ANOVA was completed. Mauchly's test indicated that the assumption of sphericity has been violated (χ^2 (2) = 11.592, p = .003), and therefore degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity (ε = .606), which showed no statistically significant difference between responses (F(1.411, 16.929) = 2.581, p = .118, η_p^2 = .177). Additionally, we investigated whether or not the trial order impacted results, using a 2-option between-subjects factor on the same rANOVA above, which revealed there was no three-way interaction between language, stimuli, and trial order (F(2, 22) = .123, p = .815, η_p^2 = .011). With these results, I selected PartWords based on their responses. In order to more strongly support learning after Experiment 2 exposure, PartWords were selected if they were rated higher or closest to the Words (see Table 7). However, for the

front language, all PartWords rated the same or lower than the Words, which may impact post-exposure familiarity ratings. The selected stimuli were grouped and tested using the same rANOVA above. With sphericity assumed (χ^2 (2) = 4.603, p = .100), the effect of language and stimuli on responses was statistically significant (F(2, 24), p = .043, η_p^2 = .231), which indicated these candidates might not be the best options. Next, we completed a pairwise comparison and discovered the statistical significance came from the relationship between back PartWords and back ShuffleWords, with a mean difference of .500, p = .024, 95% CI [0.062, 0.702]. Since the difference under investigation in the experimental procedure regards the word boundaries and the testing between Words and PartWords, we concluded this measure allows us to proceed with these stimuli.

2.4.1.4 Procedure

Experiment 2 followed the same paradigm as Experiment 1 with a few alterations. Within the background questionnaire, the question regarding handedness was removed; its inclusion is appropriate for experiments using EEG, MEG, or other methods measuring brain hemispheric performance. Next, two questions were added to the background questionnaire: one regarding a sleeping disorder diagnosis and one regarding the years of musical instruction, both of which were used as exclusionary criteria (see 2.4.1.1). Aside from updating the corresponding stimuli files and including a 5-second fade-in and -out, no other aspect of the procedure was changed.

2.4.1.5 Statistical analysis

The following analyses were performed: First, we tested for the effects of order (paused/continuous first) with a repeated measures ANOVA (rANOVA) of 2 (Condition: pause vs. continuous) x 3 (Stimulus Type: Word, PartWord, ShuffleWord) x 2 (Order: paused first vs. continuous first). Then, the effect of language was tested with an rANOVA of 2 (Condition) x 3 (Stimulus Type) x 2 (Paused Language: paused back language vs. paused front language) was completed to examine the effects of Condition, Stimulus Type, and Paused Language on familiarity ratings. Finally, an rANOVA of 2 (Condition) x 3 (Stimulus Type) was performed to test the effect of pause on familiarity

ratings. Significant interactions were investigated further with post hoc tests using a Bonferroni correction. All analyses were performed with IBM SPSS Statistics (Version 29.0.2.0).

2.4.2 Results

The rANOVA for Condition x Stimulus Type x Order revealed a main effect for Stimulus Type $F(2, 66) = 36.15$, $p < .001$, $\eta_p^2 = .523$. A Bonferroni-adjusted pairwise comparison revealed that Words were rated most familiar (M = 4.34), PartWords second-most familiar (M = 3.70), and ShuffleWords least familiar (M = 3.35), with all three types differing from each other statistically significantly ($p < .001$). There were no other statistically significant main effects. Therefore, the order did not influence results.

The rANOVA for Condition x Stimulus Type x Paused Language revealed again a main effect for Stimulus Type $F(2, 66) = 36.34$, $p < .001$, $\eta_p^2 = .524$. A Bonferroni-adjusted pairwise comparison revealed that Words were rated more familiar (M = 4.30) than PartWords (M = 3.67) and ShuffleWords (M = 3.32), all three differing with statistical significance ($p < .001$). However, there was a statistically significant interaction between Condition and Paused Language $F(1, 33) = 16.042$, $p < .001$, $\eta_p^2 = .327$. There was also a significant interaction between Condition and Stimulus Type, $F(2, 66) = 3.495$, $p = .036$, $\eta_p^2 = .096$, and a three-way interaction between Condition, Stimulus Type, and Paused Language, $F(2, 66) = 3.642$, $p = .032$, $\eta_p^2 = .099$.

Further examination into this three-way interaction between Condition, Stimulus Type, and Pause Language using Bonferroni-adjusted pairwise comparisons showed significant differences among stimuli in both languages and conditions (Figure 5). In the paused conditions: Participants showed higher familiarity to Back Words and Back PartWords, with a mean difference of 1.125, $p < .001$, 95% CI [0.590, 1.660]. Participants also showed higher familiarity for Front Words than Front PartWords, with a mean difference of 0.691, $p = .010$, 95% CI [0.141, 1.242]. In the paused condition, the difference between the Words and ShuffleWords were statistically significant for the Back ($p < .001$) and the Front ($p < .001$) languages. In the continuous conditions: There was no statistically significant difference between Front Words and Front PartWords,

with a mean difference of 0.264, p = .572, 95% CI [-0.234, 0.762], and the difference between Back Words and Back PartWords approached statistical significance, with a mean difference of 0.471, p = .081, 95% CI [-0.042, 0.983]. In the continuous condition, the difference between the Words and ShuffleWords approached significance in the Back (p = .067) and was statistically significant in the Front (p <.001) language.

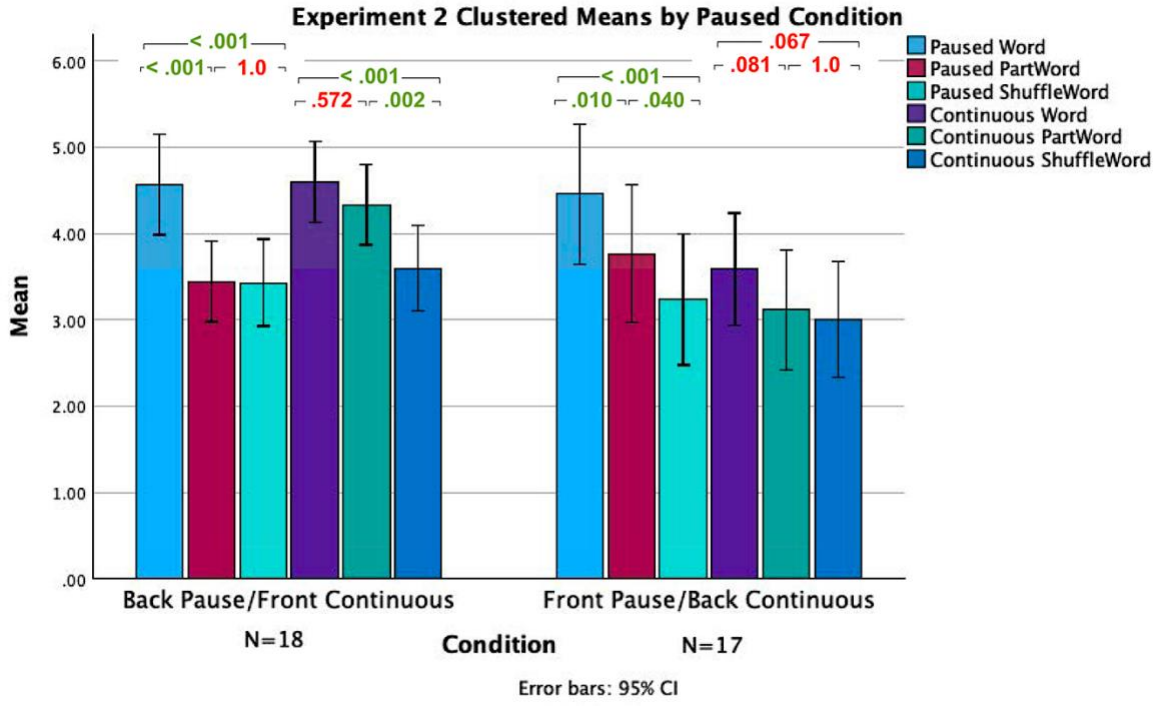

Figure 5
Experiment 2 familiarity scores categorized into languages and conditions. The Front language results are the interior six bars; the Back language results are the six exterior bars. The p-values are listed above the bars.

A final analysis combining the data from the two languages was conducted regardless of the differences presented above with the same reasoning as in Experiment 1. A 2 (Condition) x 3 (Stimulus Type) rANOVA revealed statistically significant main effect in Stimuli Type $F(2, 68) = 37.361$, p <.001, $\eta_p^2 = .524$) similarly to Experiment 1: Words were rated highest (M = 4.31), PartWords second highest (M = 3.67), and ShuffleWords third (M = 3.32). Furthermore, there was a significant interaction between Condition and Stimuli Type $F(2, 68) = 3.31$, p = .042, η_p^2 = .089. Further investigation with Bonferroni-adjusted pairwise comparisons revealed statistically significant differences in Paused Words and PartWords (p < .001) and Continuous Words and PartWords (p =

.042). These combined scores indicate support of our hypothesis that Finnish speakers can segment quadrisyllabic words within continuous speech. In addition, Words were rated different from ShuffleWords in both conditions (p <.001), and the PartWords were rated higher than ShuffleWords in the continuous (p = .006) but not in the Paused (p= .079) condition.

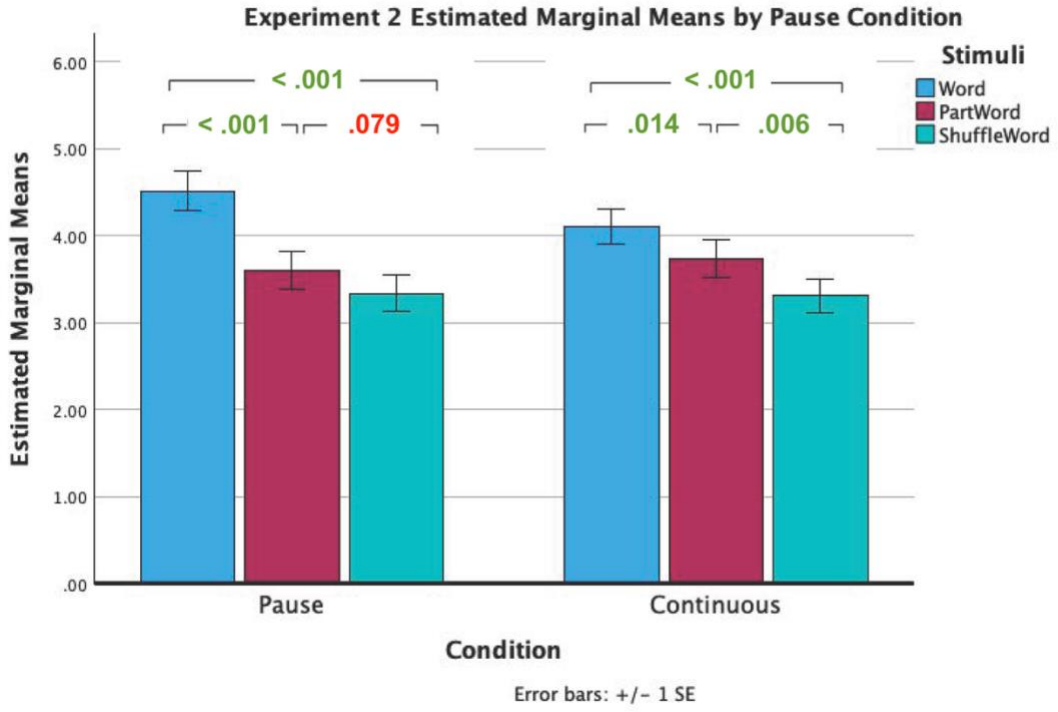

Figure 6
Experiment 2 familiarity scores distributed into pause or continuous conditions. The p-values are listed above the bars.

2.4.3 Interim discussion

The aim of Experiment 2 is to investigate the paradigm first explored in Experiment 1 with attention to stimuli creation and replication. Once the stimuli passed the confounding factors for Experiment 2, only the front language from Experiment 1 was used again, and a new back language was generated, tested for confounding factors, and pretested.

These results indicate a replication of Benjamin et al. where French-speaking adults failed to differentiate Words from PartWords in the continuous condition, suggesting

failure to segment the stream. In Experiment 2, also Finnish-speaking adults failed to segment Words from PartWords in the continuous Back (p = .081) and continuous Front (p = .572) languages. Where Benjamin et al. found successful segmentation of Words from PartWords in the paused condition (p = < .001) in French-speaking adults, we also found successful segmentation in the paused conditions (p <.001).

If the languages are separated, it shows support for Benjamin et al.'s original hypothesis that segmentation of quadrisyllabic words is only successful with the aid of a 25 millisecond pause. This also then refutes our hypothesis that word length expectations correlate to word length in word segmentation tasks, as speakers of two languages (French and Finnish) with different word length expectations (Isbilen & Christiansen, 2022; Kalimeri et al., 2015) both failed to segment a continuous stream with quadrisyllabic words. Notably, adults in both language groups have been able to segment trisyllabic words in Finnish (Kuuluvainen et al., 2025) and in French (Bagou et al., 2002; Toro et al., 2008). This indicates that word segmentation tasks may be limited by word length.

Both the current study and that of Benjamin et al (2023) suggest that prosodic cues like subliminal pauses better help participants segment words when the word length increases. Tracking TPs, which is measured by higher familiarity of Words over ShuffleWords, is not sufficient for quadrisyllabic word segmentation; the use of additional cues is needed to successfully segment words above three syllables, which is shown by the differences between Words and PartWords.

One noted objection, addressed in Benjamin et al., is the potential impact from memory capacity (2023). The Words in this paradigm are longer and therefore participants have four more syllables to memorize or learn during the study than in the trisyllabic condition. The failure to segment the speech in the continuous condition is not caused by this, however, because participants were able to segment the Words accurately with the aid of a pause. If memory capacity was indeed an issue, participants would not have been able to segment the Words in the paused condition.

3 Part III: Discussion

3.1 Experiments 1 & 2

The previous sections have outlined replication, the replication crisis, and SLL. The purpose of this thesis was to serve as a case study on replication using SLL. It explored the impact of native language word length expectations on SLL tasks and whether these expectations impacted word segmentation. Experiment 1 was a Finnish-language conceptual replication of Benjamin et al. (2023), which investigated SLL in Finnish speakers using quadrisyllabic word formations with and without pauses. During this process, we learned about SLL, word segmentation, and experimental design using an existing paradigm and experiment. Interestingly, Experiment 1 produced conflicting conclusions regarding replication: participants successfully segmented Words from PartWords in the Back language in both continuous and paused conditions, whereas the Front language was not significant in the continuous condition and merely approached significance in the paused condition. Since our languages consisted of vowel groups, these results indicate a potential vowel-placement effect on word segmentation. In this experiment, the Back language performed more like we predicted, whereas the Front language did not. Some features, including the elongated syllable “vö” in the Front language in Experiment 1, may have confounded these results; this mistake was corrected for Experiment 2. After discovering these errors, it forced us to consider new avenues for stimuli creation. Using Titone et al.’s (2024) confounding features, we discovered deeper problems with the used stimuli, such as phonological patterns in the Back language that may have aided learning during this process. It is perhaps those features that better aided word segmentation resulting in the data collected. Once those features were corrected for in Experiment 2, those stark differences seemed to dissipate.

Experiment 2 was designed to be a (mostly) direct replication of Experiment 1 by focusing on improving methodology and strengthening controls, using Titone et al.’s (2024) confounding factors. When split into languages, Experiment 2 successfully replicated Benjamin et al.’s finding that successful quadrisyllabic word segmentation is aided by a 25-millisecond pause. Experiment 2 refutes our hypothesis that Finnish speakers’ word length expectations will inform their performance on quadrisyllabic

word segmentation tasks. This suggests that these expectations do not aid segmentation in continuous speech streams with quadrisyllabic words.

The failure to find successful segmentation in the language groups in Experiment 2 indicates that native language word length expectations, as we initially hypothesized, may not impact word segmentation in the absence of pauses. This would evidently support Benjamin et al.'s original claim. However, due to the weak statistical power in this analysis, we need more participants to support our conclusion.

Interestingly, the performance of language groups differed in each experiment. Both languages (front-vowel or back-vowel) performed slightly different from one another. This could indicate that vowel placement may impact performance on statistical language learning paradigms. Are front vowels more difficult to process than back vowels? Do Finnish speakers process these vowels differently based on the language's reliance on these differences?

Ultimately, what this illustrates is that the languages used in a word segmentation paradigm such as this, especially when the languages are meant to be collated for analysis, the languages must perform similarly during testing. Using proper stimuli is dependent upon how similarly participants respond to them. Experiment 1 showed that the pairing of those languages was too different to produce cohesive results. When improved for Experiment 2, however, that difference was mitigated, albeit not entirely.

Once the data is analyzed together to improve statistical power, however, the story changes back to support our main hypothesis. The differences between groups dissipates and both the continuous and paused conditions show participants can segment Words from PartWords in both conditions. The reason we combined the scores this way is because the statistical power is weakened once these groups are distributed into language groups; we need more participants in each group to confidently conclude results.

In conclusion, since the combined results hold higher statistical power, these results support our hypothesis that Finnish speakers have higher word length expectations and can segment quadrisyllabic word formations in continuous speech. This supports that

native language may influence word length expectations and thus impact our ability to segment speech in a learning environment.

3.2 Student replication

The aim of this project was to illustrate the benefits of student replication. Both experiments highlight the difficulties in producing quality student work (see the limitations in both experiments) and contributing to the scientific conversation at the same time. While teaching is an important endeavor in forming the next generation of scientists, there may be a danger in utilizing student work in direct conversation with original studies, simply because the lack of a rigorous peer review process. The review process for a thesis differs across universities and across countries; but if we want to consider something scientific, it must undergo some level of peer review. If we conflate that master's student's works are an extension of the researcher's work, and that any researcher's work should be investigated by another researcher before publication, should others from different universities or faculties also review master's student's work in order to consider it peer reviewed? If it is not considered peer review work, does that necessitate the thesis's absence from scientific discourse? And does any student work need to be peer reviewed? Although I agree that the review process looks different for student work (if it exists at all), student replication can still be considered supplementary material for an original study. It would indicate not only interest in the topic investigated, but it may also provide valuable information to the field. By directly indicating a study was completed by students, it would alert readers to the weak form of peer review during that study; instead of concretely claiming successful or failed replication (by the nature of not going through a peer review process), it would provide a gauge of replicability for others to investigate more deeply. In this way, student work would be a valuable learning tool but also can inform other researchers about replicability in their field.

This thesis serves as a practical case study of student-led replication attempts, which supports the suggestion that PhD candidates (and by extension, master's students) can contribute to their fields and improve the state of replicability. This can be done through

two avenues: utilizing group work for course credit (Experiment 1) or through individual student work like a thesis (Experiment 2).

Group assignments can produce valuable insight into replication, even if the assignments themselves produce conceptual replications. That work, which functions to introduce many aspects of the scientific process to students (such as background literature, experimental design, methodologies, and analysis), will also produce valuable knowledge to the corresponding field. By allowing students the opportunity to follow the workflow of an established study, students not only receive a better image of the experience behind experimentation, but it also allows students to learn about scientific processes, produce valuable knowledge to the field, and become aware of the importance of replication through practice.

Experiment 2 highlights the second approach to student replication work, where students can acquaint themselves to the individual practice of the scientific process and the personal actions that influence the replication crisis (e.g., QRPs and statistical analyses). This stage allows the student to follow the guidelines from an original study (again the same benefits from group assignments), but this allows greater attention toward the practice of working as a researcher. In this way, guidance comes from more than the student's immediate supervisor; it also comes from the original authors of the study via their published article. This work is an intense independent study, while attempting to replicate results. It allows students to independently participate in open science, such as filing preregistrations and making data and methods available.

The overall approach to student replication work is not to produce novel studies; it is to teach and to produce knowledge in the field. Teaching is through an apprentice-type work, using the original study as a learning tool and guide. Producing knowledge is by contributing valuable information in support or contesting the original study. However, by removing the notion of novelty from the student's attention (and thus targeting neophilia), it allows a tailored focus on the practice of scientific research and producing quality over novelty.

This thesis highlights some of the challenges and learning opportunities inherent in replication. After first attempting a conceptual replication and then a more direct

replication, this thesis focused on refining methods. This kind of improvement is crucial given the concerns of the replication crisis across disciplines. The emphasis on stimuli creation and addressing the confounding factors in Experiment 2 aligns with the attention given to methodology and statistical analysis as part of the solution to the crisis. Transparency has been demonstrated throughout this process, following the principles of open science, including preregistration. Attempting replication, even with limitations, contributes to the health check-up that replication may provide, offering valuable insight into the state of replicability within SLL research.

3.3 Limitations

Limitations remained throughout these studies, especially when it comes to the stimuli used. Although significant effort was placed on controlling the factors of stimuli, challenges remained. In Experiment 2, one Word still included an exclusionary sound "-na." When selecting stimuli based on the targeted range of positional frequencies (20–45%), challenges remained for the best fit. Several candidates remained outside the targeted range, and this could have impacted results accordingly. This constraint could have been overlooked had more time been available to iterate more alternative stimuli. An error was also present in Experiment 1's stimuli, with the lengthening of "vö" in the audio file. The potential of phonological rhythms impacting results still remains. Although they were better controlled for in Experiment 2, the risk for impact remains.

Another limitation is the lower number of participants in both experiments (38 in Experiment 1, 35 in Experiment 2, after exclusions) resulting in low statistical power when divided into the corresponding language groups, which necessitates replication with larger sample sizes.

Another limitation, unfortunately, was that no attempt to contact the original authors was made. However, according to Kahneman, an original author should have been contacted at the beginning of the replication study (2014). By the nature of a master's thesis, I concluded that the attention given to such a small project would not warrant contact. Had this project proceeded with more direct replication of Benjamin et al., contacting one original author would have been warranted. To accommodate such a

strong claim that replication was not successfully completed, I completed a replication of our study to verify results, which is a recommended action Kahneman suggests as well (Bohannon, 2014). It is my suggestion for those students aiming to replicate, however, that contact should be made. This is not to only accommodate replication etiquette; it is to introduce students to a more collaborative research environment.

As a master's student, executing mastery of every aspect considered here may prove challenging: completing preregistration, accommodating replication etiquette, minimizing QRPs, completing appropriate statistical analyses, and executing experimental design, methods, and analyses, all while learning a (potentially) novel topic. This level of difficulty, which may be a graduate-level expectation, also leaves the potential for error at every instance. This kind of student error, however, is perhaps the biggest limitation on Perrault's suggestion of including student work in conversation with an original study. Student work goes through a small level of peer review (e.g., feedback from advisors, feedback after submission of a master's thesis), but this low level may be unsatisfactory for the larger scientific community. Theses do not always iterate through multiple rounds of review before final submission. However, labeling each study as "student work" would already indicate to readers of this low-level peer-review status and allow students to remain part of the conversation. For instance, with all the limitations outlined herein, this thesis could serve as an example for future students to produce replication attempts.

3.4 Future research

Future studies in SLL could employ these methods outlined by Titone et al. (2024), carefully addressing each confounding factor when generating new stimuli. For trisyllabic word formation tasks, their set of codes may streamline the process of stimuli creation. Attention to stimuli is also important considering a potential impact from vowel placement. Further work should investigate the relationship between vowel placement (perhaps using Finnish's strict vowel harmony) during word segmentation tasks. This could look into whether this distinction is seen only in quadrisyllabic formations or in the trisyllabic ones as well.

Other work should investigate quadrisyllabic word segmentation in other languages of varying lengths. Having French and Finnish participants together in the same study would help to illustrate this behavior better, perhaps controlling more for the language similarities as to not confound results.

Digit span tests, which may be used to measure working memory, were collected from participants, and those measures could explore the relationship between working and familiarity ratings in word segmentation tasks. Using those scores in relationship to their familiarity ratings would prove insightful if not novel.

Student replication projects should be integrated as part of graduate training, which provides practical research experience and contributes valuable data to the scientific community. By fulfilling this request, students produce work that tackles the replication crisis and allows learning opportunities. Developing a database or repository for student replication work could improve communication and accessibility, especially if the original study could direct toward the student work, indicating to readers that a student replication study was completed (Balafoutas et al., 2025).

As always, a replication of Experiment 2 should be completed with a larger sample size to increase statistical power and confirm the findings for the necessity of pauses in quadrisyllabic word segmentation.

Acknowledgements

A very special thanks to Soila Kuuluvainen for the dedicated teaching and supervision over the last two years. This work could not have been completed without her continued guidance, and I am sincerely grateful for everything she has taught me. Thank you to Saara Kaskivuo for providing the Python scripts that got me started on all my stimuli analyses. Those scripts made me excited to work on this project. Thank you to Martti Vainio for providing audio files for this and many other projects. Thanks to Riikka Möttönen for her supervision and feedback in research group meetings and thesis discussion groups. Thanks to the work by Benjamin et al. (2023) and Titone et al. (2024); these two works are foundational to the work above, and I'm thankful to have used their works as tools and guides. Thanks to Freja Kivikettu, Melissa Adkins-Hempel, Nathan Knoblauch, Benjamin Brodie, and others for the continued support throughout this project.

References

Antonakis, J. (2017). On doing better science: From thrill of discovery to policy implications. *The Leadership Quarterly*, *28*(1), 5–21. https://doi.org/10.1016/j.leaqua.2017.01.006

Arpinon, T., & Espinosa, R. (2023). A practical guide to Registered Reports for economists. *Journal of the Economic Science Association*, *9*(1), 90–122. https://doi.org/10.1007/s40881-022-00123-1

Artner, R., Verliefde, T., Steegen, S., Gomes, S., Traets, F., Tuerlinckx, F., & Vanpaemel, W. (2021). The reproducibility of statistical results in psychological research: An investigation using unpublished raw data. *Psychological Methods*, *26*(5), 527–546. https://doi.org/10.1037/met0000365

Bagou, O., Fougeron, C., & Frauenfelder, U. H. (2002). Contribution of prosody to the segmentation and storage of "words" in the acquisition of a new mini-language. *Speech Prosody 2002*, 159–162. https://doi.org/10.21437/SpeechProsody.2002-25

Bahri, M., Farrahi, H., Mahdavinataj, H., & Batouli, S. A. H. (2024). Eight brain structures mediate the age-related alterations of the working memory: Forward and backward digit span tasks. *Frontiers in Psychology*, *15*, 1377342. https://doi.org/10.3389/fpsyg.2024.1377342

Balafoutas, L., Celse, J., Karakostas, A., & Umashev, N. (2025). Incentives and the replication crisis in social sciences: A critical review of open science practices. *Journal of Behavioral and Experimental Economics*, *114*, 102327. https://doi.org/10.1016/j.socec.2024.102327

Benjamin, L., Fló, A., Palu, M., Naik, S., Melloni, L., & Dehaene-Lambertz, G. (2023). Tracking transitional probabilities and segmenting auditory sequences are dissociable processes in adults and neonates. *Developmental Science*, *26*(2), e13300. https://doi.org/10.1111/desc.13300

Berg, T., Zörnig, P., & Lehr, C. (2022). The effects of type and token frequency on word length: A cross-linguistic study. *Glottotheory*, *13*(2), 173–209. https://doi.org/10.1515/glot-2022-2007

Besner, D., & Smith, M. C. (1992). Chapter 3 Basic Processes in Reading: Is the Orthographic Depth Hypothesis Sinking? In R. Frost & L. Katz (Eds.), *Advances in Psychology* (Vol. 94, pp. 45–66). North-Holland. https://doi.org/10.1016/S0166-4115(08)62788-0

Black, A., & Bergmann, C. (2017). Quantifying Infants' Statistical Word Segmentation: A Meta-Analysis. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (pp. 124–129).

Boersma, P., & Weenink, D. (2025). *Praat: Doing phonetics by computer* (Version 6.4.27) [Computer software]. http://www.praat.org/

Bohannon, J. (2014). Replication effort provokes praise—And 'bullying' charges. *Science*, *344*(6186), 788–789. https://doi.org/10.1126/science.344.6186.788

Brodeur, A., Kattan, L., & Musumeci, M. (2024). *Job Market Stars* (GLO Discussion Paper No. 1514; Global Labor Organization (GLO)).

Buiatti, M., Pena, M., & Dehaenelambertz, G. (2009). Investigating the neural correlates of continuous speech computation with frequency-tagged neuroelectric responses. *NeuroImage*, *44*(2), 509–519. https://doi.org/10.1016/j.neuroimage.2008.09.015

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., & Wu, H. (2016). Evaluating replicability of

laboratory experiments in economics. *Science*, *351*(6280), 1433–1436. https://doi.org/10.1126/science.aaf0918

Carney, D. R., Cuddy, A. J. C., & Yap, A. J. (2010). Power Posing: Brief Nonverbal Displays Affect Neuroendocrine Levels and Risk Tolerance. *Psychological Science*, *21*(10), 1363–1368. https://doi.org/10.1177/0956797610383437

Chambers, C. D., & Tzavella, L. (2021). The past, present and future of Registered Reports. *Nature Human Behaviour*, *6*(1), 29–42. https://doi.org/10.1038/s41562-021-01193-7

Dalton, D. R., Aguinis, H., Dalton, C. M., Bosco, F. A., & Pierce, C. A. (2012). REVISITING THE FILE DRAWER PROBLEM IN META-ANALYSIS: AN ASSESSMENT OF PUBLISHED AND NONPUBLISHED CORRELATION MATRICES. *Personnel Psychology*, *65*(2), 221–249. https://doi.org/10.1111/j.1744-6570.2012.01243.x

Derksen, M. (2021). A Menagerie of Imposters and Truth-Tellers:Diederik Stapel and the Crisis in Psychology. In S. Woolgar, E. Vogel, D. Moats, & C.-F. Helgesson (Eds.), *The Imposter as Social Theory: Thinking with Gatecrashers, Cheats and Charlatans* (pp. 53–75). Bristol University Press. https://www.jstor.org/stable/j.ctv1p6hphs.8

Derksen, M., & Field, S. (2022). The Tone Debate: Knowledge, Self, and Social Order. *Review of General Psychology*, *26*(2), 172–183. https://doi.org/10.1177/10892680211015636

Diener, E., & Biswas-Diener, R. (2025). *The Replication Crisis in Psychology*. Noba. https://nobaproject.com/modules/the-replication-crisis-in-psychology

Ferraro, P. J., & Shukla, P. (2023). Credibility crisis in agricultural economics. *Applied Economic Perspectives and Policy*, *45*(3), 1275–1291. https://doi.org/10.1002/aepp.13323

Field, S., Van Dongen, N., & Tiokhin, L. (2024). Reflections on the Unintended Consequences of the Science Reform Movement. *Journal of Trial and Error*, *4*(1), 1–4. https://doi.org/10.36850/ed4

Fiser, J., & Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences*, *99*(24), 15822–15826. https://doi.org/10.1073/pnas.232472899

Fiske, A. P. P. S. T. (2016). A Call to Change Science's Culture of Shaming. *APS Observer*, *29*. https://www.psychologicalscience.org/observer/a-call-to-change-sciences-culture-of-shaming

Franco, A., Malhotra, N., & Simonovits, G. (2016). Underreporting in Psychology Experiments: Evidence From a Study Registry. *Social Psychological and Personality Science*, *7*(1), 8–12. https://doi.org/10.1177/1948550615598377

Franklin, A. (2018). *Is It the 'Same' Result: Replication in Physics:* Morgan & Claypool Publishers. https://doi.org/10.1088/978-1-64327-162-0

Frias-Navarro, D., Pascual-Llobell, J., Pascual-Soler, M., Perezgonzalez, J., & Berrios-Riquelme, J. (2020). Replication crisis or an opportunity to improve scientific production? *European Journal of Education*, *55*(4), 618–631. https://doi.org/10.1111/ejed.12417

Gabelica, M., Bojčić, R., & Puljak, L. (2022). Many researchers were not compliant with their published data sharing statement: A mixed-methods study. *Journal of Clinical Epidemiology*, *150*, 33–41. https://doi.org/10.1016/j.jclinepi.2022.05.019

Goldenberg, M. J. (2023). Public trust in science. *Interdisciplinary Science Reviews*, *48*(2), 366–378. https://doi.org/10.1080/03080188.2022.2152243

Gómez Varela, I., Orpella, J., Poeppel, D., Ripolles, P., & Assaneo, M. F. (2024). Syllabic rhythm and prior linguistic knowledge interact with individual differences to modulate phonological statistical learning. *Cognition*, *245*, 105737. https://doi.org/10.1016/j.cognition.2024.105737

Gorilla Experiment Builder. (n.d.). Retrieved February 24, 2025, from https://gorilla.sc/

Haig, B. D. (2022). Understanding Replication in a Way That Is True to Science. *Review of General Psychology*, *26*(2), 224–240. https://doi.org/10.1177/10892680211046514

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The Extent and Consequences of P-Hacking in Science. *PLOS Biology*, *13*(3), e1002106. https://doi.org/10.1371/journal.pbio.1002106

Hendrick, C. (1991). Replication, Strict Replications, and Conceptual Replications: Are They Important? In J. Neuliep W. (Ed.), *Replication Research in the Social Sciences* (pp. 41–49). Sage Publications.

Huntington-Klein, N., Arenas, A., Beam, E., Bertoni, M., Bloem, J. R., Burli, P., Chen, N., Grieco, P., Ekpe, G., Pugatch, T., Saavedra, M., & Stopnitzky, Y. (2021). The influence of hidden researcher decisions in applied microeconomics. *Economic Inquiry*, *59*(3), 944–960. https://doi.org/10.1111/ecin.12992

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Isbilen, E. S., & Christiansen, M. H. (2022). Statistical Learning of Language: A Meta-Analysis Into 25 Years of Research. *Cognitive Science*, *46*(9), e13198. https://doi.org/10.1111/cogs.13198

Jasny, B. R., Chin, G., Chong, L., & Vignieri, S. (2011). Again, and Again, and Again …. *Science*, *334*(6060), 1225. https://doi.org/10.1126/science.334.6060.1225

Johnson, D. J., Cheung, F., & Donnellan, M. B. (2014). Does Cleanliness Influence Moral Judgments?: A Direct Replication of Schnall, Benton, and Harvey (2008). *Social Psychology*, *45*(3), 209–215. https://doi.org/10.1027/1864-9335/a000186

Jones, G., & Macken, B. (2015). Questioning short-term memory and its measurement: Why digit span measures long-term associative learning. *Cognition*, *144*, 1–13. https://doi.org/10.1016/j.cognition.2015.07.009

Kahneman, D. (2014). A New Etiquette for Replication. *Social Psychology*, *45*(4), 299–311.

Kalimeri, M., Constantoudis, V., Papadimitriou, C., Karamanos, K., Diakonos, F. K., & Papageorgiou, H. (2015). Word-length Entropies and Correlations of Natural Language

Written Texts. *Journal of Quantitative Linguistics*, *22*(2), 101–118. https://doi.org/10.1080/09296174.2014.1001636

Keating, D. M., & Totzkay, D. (2019). We do publish (conceptual) replications (sometimes): Publication trends in communication science, 2007–2016. *Annals of the International Communication Association*, *43*(3), 225–239. https://doi.org/10.1080/23808985.2019.1632218

Kelih, E. (2012). Systematic Interrelations Between Grapheme Frequencies and Word Length: Empirical Evidence from Slovene*. *Journal of Quantitative Linguistics*, *19*(3), 205–231. https://doi.org/10.1080/09296174.2012.685304

Keppel, G. (1982). *Design and Analysis: A Researcher's Handbook* (Second). Prentice-Hall.

Kiai, A., & Melloni, L. (2021). *What canonical online and offline measures of statistical learning can and cannot tell us*. https://doi.org/10.1101/2021.04.19.440449

Kohrt, F., Smaldino, P. E., McElreath, R., & Schönbrodt, F. (2023). *The coevolution of effort and replication, recreated, replicated and corrected* [Graphic]. Zenodo. https://doi.org/10.5281/ZENODO.7547729

Kuuluvainen, S., Kaskivuo, S., Vainio, M., Smalle, E., & Möttönen, R. (2025). Prosody enhances learning of statistical dependencies from continuous speech streams in adults. *Cognition*, *262*, 106169. https://doi.org/10.1016/j.cognition.2025.106169

Lykken, D., T. (1968). Statistical Significance in Psychological Research. *Psychological Bulletin*, *70*(3p1), 151–159. https://doi.org/10.1037/h0026141

Mandikal Vasuki, P. R., Sharma, M., Demuth, K., & Arciuli, J. (2016). Musicians' edge: A comparison of auditory processing, cognitive abilities and statistical learning. *Hearing Research*, *342*, 112–123. https://doi.org/10.1016/j.heares.2016.10.008

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of *p* values just below .05. *Quarterly Journal of Experimental Psychology*, *65*(11), 2271–2279. https://doi.org/10.1080/17470218.2012.711335

Mattys, S. L., White, L., & Melhorn, J. F. (2005). Integration of Multiple Speech Segmentation Cues: A Hierarchical Framework. *Journal of Experimental Psychology: General*, *134*(4), 477–500. https://doi.org/10.1037/0096-3445.134.4.477

McCrary, J., Christensen, G., & Fanelli, D. (2016). Conservative Tests under Satisficing Models of Publication Bias. *PLOS ONE*, *11*(2), e0149590. https://doi.org/10.1371/journal.pone.0149590

Miguel, E. (2021). Evidence on Research Transparency in Economics. *Journal of Economic Perspectives*, *35*(3), 193–214. https://doi.org/10.1257/jep.35.3.193

Miłkowski, M., Hensel, W. M., & Hohol, M. (2018). Replicability or reproducibility? On the replication crisis in computational neuroscience and sharing only relevant detail. *Journal of Computational Neuroscience*, *45*(3), 163–172. https://doi.org/10.1007/s10827-018-0702-z

Mollayeva, T., Thurairajah, P., Burton, K., Mollayeva, S., Shapiro, C. M., & Colantonio, A. (2016). The Pittsburgh sleep quality index as a screening tool for sleep dysfunction in clinical and non-clinical samples: A systematic review and meta-analysis. *Sleep Medicine Reviews*, *25*, 52–73. https://doi.org/10.1016/j.smrv.2015.01.009

Musich, M., Beversdorf, D. Q., McCrae, C. S., & Curtis, A. F. (2024). Subjective–Objective Sleep Discrepancy in a Predominately White and Educated Older Adult Population: Examining the Associations With Cognition and Insomnia. *The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences*, *79*(7), gbae074. https://doi.org/10.1093/geronb/gbae074

National Academies of Sciences, E., Affairs, P. and G., Committee on Science, E., Information, B. on R. D. and, Sciences, D. on E. and P., Statistics, C. on A. and T., Analytics, B. on M. S. and, Studies, D. on E. and L., Board, N. and R. S., Education, D. of B. and S. S. and, Statistics, C. on N., Board on Behavioral, C., & Science, C. on R. and R. in. (2019). Understanding Reproducibility and Replicability. In *Reproducibility and Replicability in*

Science. National Academies Press (US). https://www.ncbi.nlm.nih.gov/books/NBK547546/

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Struhl, M. K., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, Robustness, and Reproducibility in Psychological Science. *Annual Review of Psychology*, *73*(1), 719–748. https://doi.org/10.1146/annurev-psych-020821-114157

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. https://doi.org/10.1177/1745691612459058

Olsthoorn, N. M., Andringa, S., & Hulstijn, J. H. (2014). Visual and auditory digit-span performance in native and non-native speakers. *International Journal of Bilingualism*, *18*(6), 663–673. https://doi.org/10.1177/1367006912466314

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Ordin, M., Polyanskaya, L., Laka, I., & Nespor, M. (2017). Cross-linguistic differences in the use of durational cues for the segmentation of a novel language. *Memory & Cognition*, *45*(5), 863–876. https://doi.org/10.3758/s13421-017-0700-9

Pashler, H., & Wagenmakers, E. (2012). Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. https://doi.org/10.1177/1745691612465253

Pavlovia. (n.d.). Pavlovia. https://pavlovia.org

Peña, M., Bonatti, L. L., Nespor, M., & Mehler, J. (2002). Signal-Driven Computations in Speech Processing. *Science*, *298*(5593), 604–607. https://doi.org/10.1126/science.1072901

Perrault, E. K. (2023). Teaching replication through replication to solve the replication "crisis." *Communication Teacher*, *37*(3), 220–226. https://doi.org/10.1080/17404622.2022.2123110

Peterson, D. (2021). The Replication Crisis Needs Field-Specific Solutions. *Nature*, *594*, 151.

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, *10*(9), 712–712. https://doi.org/10.1038/nrd3439-c1

Ranehill, E., Dreber, A., Johannesson, M., Leiberg, S., Sul, S., & Weber, R. A. (2015). Assessing the Robustness of Power Posing: No Effect on Hormones and Risk Tolerance in a Large Sample of Men and Women. *Psychological Science*, *26*(5), 653–656. https://doi.org/10.1177/0956797614553946

Riemann, D., Spiegelhalder, K., Feige, B., Voderholzer, U., Berger, M., Perlis, M., & Nissen, C. (2010). The hyperarousal model of insomnia: A review of the concept and its evidence. *Sleep Medicine Reviews*, *14*(1), 19–31. https://doi.org/10.1016/j.smrv.2009.04.002

Rosenthal, R. (1979). The "File Drawer Problem" and Tolerance for Null Results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Rosenthal, R., & Rosnow, R. L. (2008). *Essentials of Behavioral Research: Methods and Data Analysis* (3rd ed.). McGraw-Hill.

Ruggles, D. R., Freyman, R. L., & Oxenham, A. J. (2014). Influence of Musical Training on Understanding Voiced and Whispered Speech in Noise. *PLoS ONE*, *9*(1), e86980. https://doi.org/10.1371/journal.pone.0086980

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. *Science*, *274*(5294), 1926–1928. 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. 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. https://doi.org/10.1006/jmla.1996.0032

Sarewitz, D. (2012). Beware the creeping cracks of bias. *Nature*, *485*(7397), 149–149. https://doi.org/10.1038/485149a

Sargent, C. L. (1981). The Repeatability of Significance and the Significance of Repeatability. *European Journal of Parapsychology*, *3*(4), 423–443.

Scheel, A. M., Schijen, M. R. M. J., & Lakens, D. (2021). An Excess of Positive Results: Comparing the Standard Psychology Literature With Registered Reports. *Advances in Methods and Practices in Psychological Science*, *4*(2), 1–12. https://doi.org/10.1177/2515245921100

Schmidt, S. (2009). Shall we Really do it Again? The Powerful Concept of Replication is Neglected in the Social Sciences. *Review of General Psychology*, *13*(2), 90–100. https://doi.org/10.1037/a0015108

Schnall, S. (2014a, May 22). *An Experience with a Registered Replication Project* [Text]. https://www.psychol.cam.ac.uk/cece/blog

Schnall, S. (2014b, November 18). *Social Media and the Crowd-Sourcing of Social Psychology* [Text]. https://www.psychol.cam.ac.uk/cece/blog

Schnall, S., Benton, J., & Harvey, S. (2008). With a Clean Conscience: Cleanliness Reduces the Severity of Moral Judgments. *Psychological Science*, *19*(12), 1219–1222. https://doi.org/10.1111/j.1467-9280.2008.02227.x

Schön, D., Boyer, M., Moreno, S., Besson, M., Peretz, I., & Kolinsky, R. (2008). Songs as an aid for language acquisition. *Cognition*, *106*(2), 975–983. https://doi.org/10.1016/j.cognition.2007.03.005

Schroeder, R. W., Twumasi-Ankrah, P., Baade, L. E., & Marshall, P. S. (2012). Reliable Digit Span: A Systematic Review and Cross-Validation Study. *Assessment*, *19*(1), 21–30. https://doi.org/10.1177/1073191111428764

Serra-Garcia, M., & Gneezy, U. (2021). Nonreplicable publications are cited more than replicable ones. *Science Advances*, *7*(21), eabd1705. https://doi.org/10.1126/sciadv.abd1705

Shadish, W., Cook, T., & Campbell, D. (2002). Quasi-Experiments: Interrupted Time-Series Designs. In *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (pp. 171–206). Houghton Mifflin.

Simons, D. J. (2014). The Value of Direct Replication. *Perspectives on Psychological Science*, *9*(1), 76–80. https://doi.org/10.1177/1745691613514755

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), 160384. https://doi.org/10.1098/rsos.160384

Soares, A. P., França, T., Gutiérrez-Domínguez, F.-J., Sousa, I., & Oliveira, H. M. (2023). As trials go by: Effects of 2-AFC item repetition on statistical learning performance. *Canadian Journal of Experimental Psychology / Revue Canadienne de Psychologie Expérimentale*, *77*(1), 57–72. https://doi.org/10.1037/cep0000290

Stevens, J. R. (2017). Replicability and Reproducibility in Comparative Psychology. *Frontiers in Psychology*, *8*, 862. https://doi.org/10.3389/fpsyg.2017.00862

Strait, D. L., Kraus, N., Parbery-Clark, A., & Ashley, R. (2010). Musical experience shapes top-down auditory mechanisms: Evidence from masking and auditory attention performance. *Hearing Research*, *261*(1–2), 22–29. https://doi.org/10.1016/j.heares.2009.12.021

Suomi, K., Mcqueen, J. M., & Cutler, A. (1997). Vowel Harmony and Speech Segmentation in Finnish. *Journal of Memory and Language*, *36*(3), 422–444. https://doi.org/10.1006/jmla.1996.2495

Titone, L., Milosevic, N., & Meyer, L. (2024). *The ARC Toolbox: Artificial Languages with Rhythmicity Control*. https://doi.org/10.1101/2024.05.24.595268

Tomczak, J., Gordon, A., Adams, J., Pickering, J. S., Hodges, N., & Evershed, J. K. (2023). What over 1,000,000 participants tell us about online research protocols. *Frontiers in Human Neuroscience*, *17*, 1228365. https://doi.org/10.3389/fnhum.2023.1228365

Toro, J. M., Nespor, M., Mehler, J., & Bonatti, L. L. (2008). Finding Words and Rules in a Speech Stream: Functional Differences Between Vowels and Consonants. *Psychological Science*, *19*(2), 137–144. https://doi.org/10.1111/j.1467-9280.2008.02059.x

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An Agenda for Purely Confirmatory Research. *Perspectives on Psychological Science*, *7*(6), 632–638. https://doi.org/10.1177/1745691612463078

West, J. D., & Bergstrom, C. T. (2021). Misinformation in and about science. *Proceedings of the National Academy of Sciences*, *118*(15), e1912444117. https://doi.org/10.1073/pnas.1912444117

Wilholt, T. (2013). Epistemic Trust in Science. *The British Journal for the Philosophy of Science*, *64*(2), 233–253. https://doi.org/10.1093/bjps/axs007