Seeking reproducibility: Assessing a multimodal study of the testing effect

Marc Beardsley, Davinia Hernández-Leo, Rafael Ramirez-Melendez

Universitat Pompeu Fabra, Barcelona

Practitioner Notes

What is already known about this subject matter?

- Multimodal data provides a more complete picture of behavioural effects.
- Low-cost devices are enabling the greater use of multimodal data.
- Multimodal data requires advanced methods of collecting and combining data.
- Reproducibility is a key tenet of science, yet many studies are not easily reproducible.

What this paper adds?

- An exploration of the complexity multimodal data adds to reproducibility.
- A replication study of the behavioural and physiological effects of testing versus restudying.
- A critical analysis of the challenges faced in reproducing a multimodal study.

Implications for practitioners

- Reproducible studies need to be effectively planned from the start.
- Reproducibility is likely to become a requirement for publication.
- Multimodal data makes reproducibility more challenging.

Abstract

Low-cost devices have widened the use of multimodal data in experiments providing a more complete picture of behavioural effects. However, the accurate collection and combination of multimodal and behavioural data in a manner that enables reproducibility is challenging and often requires researchers to refine their approaches. This paper presents a direct replication of a multimodal wordlist experiment. Specifically, we use a low-cost Emotiv EPOC® to acquire electrophysiological measures of brain activity to investigate whether retrieval during learning facilitates the encoding of subsequent learning as measured by performance on recall tests and reflected by changes in alpha wave oscillations. Behavioural results of the wordlist experiment were replicated but physiological results were not. We conclude the paper by highlighting the challenges faced in terms of replicating the previous work and in attempting to facilitate the reproducibility of our own experiment.

Keywords: reproducibility, multimodal data, EEG, testing effect

Introduction

Advances in technology and our understanding of the cognitive processes that underlie learning (Baddeley, 2012; Van Merrienboer & Sweller, 2005) enable a greater number of researchers to examine educational approaches from both behavioural and physiological perspectives to better understand the observed effects in learners. Our aim is to investigate whether a simple alteration to the presentation of study material could both improve long-term retention of learned material and enhance encoding of subsequently presented new material in educational settings as demonstrated behaviourally by performances on recall tests and physiologically by patterns of oscillatory brain activity in the alpha frequency range as measured by an electroencephalogram (EEG). Alpha-band oscillations "play an active role in information processing" (Klimesh, 2012) and "oscillatory alpha power recorded with magneto- or electroencephalography (M/EEG; 8–13 Hz) is studied extensively in the fields of attention and working memory" (Wilsch et al., 2014).

Our study attempts to replicate a wordlist experiment of Pastötter et al. (2011) while adopting a more accessible approach that makes use of a low-cost device, open source software, and conducts the experiment in a non-laboratory setting. The purpose of the study is to validate an accessible, multimodal setup upon which future conceptual replications of the wordlist experiment and the testing effect may be conducted as new competencies are often required in multimodal studies to effectively make use of specialized equipment (Schmidt, 2009); apply appropriate methods of data acquisition, cleaning, analysis and reporting (Carp, 2012); and adequately facilitate reproducibility through a transparent documentation of methods and data analysis including the making of methods, data, code and workflows openly available (Stodden et al., 2016).

The simple alteration to the presentation of study material to learners being investigated is the insertion of retrieval activities in the form of tests. Test-enhanced learning is an approach based on "*the testing effect*" which studies have shown enhances long-term retention via retrieval practice (Roediger & Karpicke, 2006). The approach can be incorporated into the presentation of learning materials by educators or adopted directly by students independent of the learning environment as an effective self-study technique (Bjork et al., 2013). Through the replication attempt, this paper looks at the former as it attempts to replicate the effects of interpolated tests on wordlist recall with a focus on the forward effect of testing. The more studied backward effect of testing states that testing enhances the retention of previously presented material whereas the forward effect of testing suggests that testing enhances the encoding of subsequently presented material (Pastötter & Bäuml, 2014; Tulving & Watkins, 1974). The paper contributes our replication attempt, details challenges involved in completing robust interdisciplinary work, and highlights efforts taken to improve the usefulness of our contribution to scientific knowledge by better aligning our multimodal study to reproducibility principles (Button et al., 2013; Nosek et al., 2012).

Reproducibility in Science

Researchers work collectively to contribute to the construction and validation of scientific knowledge. The Open Science Collaboration (2015) writes that "scientific claims should not gain credence because of the status or authority of their originator but by the replicability of their supporting evidence" as replications offer proof that an experiment's findings can be extended beyond specific contextual circumstances and reflect a broader knowledge (Schmidt, 2009). Button et al., (2013) also highlight the importance of reproducibility

in generating reliable research and curtly state that "unreliable research is inefficient and wasteful."

The current state

Button et al., (2013) declare that there is "substantial evidence that a large proportion of the evidence reported in the scientific literature may be unreliable." This claim is supported by their analysis of 48 neuroscientific articles which included data from 49 meta-analyses and 730 individual primary studies that found the median statistical power of the neuroscientific studies reviewed to be 21%. Along similar lines, the Open Science Collaboration (2015) conducted replications of psychology studies that had been published in prominent psychology journals and were only able to reproduce the findings in 37 of the 97 published studies that reported significant results. The above studies coincide with the results of a survey in *Nature* (Baker, 2016) in which more than 70% of the 1,576 responding researchers confirmed that they had tried and failed to reproduce another scientist's experiments, more than 50% failed to reproduce their own experiments, and 52% agreed with the statement that there is a significant 'crisis' of reproducibility. Finally, Ioannidis (2005) demonstrated through simulations that for most study designs it can be "proven that most claimed research findings are false" due to factors such as researcher flexibility, biases, small sample sizes, small effect sizes, and a lack of consideration for the totality of the evidence partially demonstrated by a failing to account for prior probabilities of true to no relationships.

Contributing factors

It has been suggested that the underwhelming state of reproducibility in science is a result of the publication incentive system teamed with the degrees of freedom researchers have both in terms of the methods and analysis applied and in what level of details they are reported (Munafò

et al., 2017; Nosek et al., 2012; Simmons et al., 2011). Furthermore, an over reliance on statistical significance by publications works to bias the decisions researchers make with their degrees of freedom. As Munafò et al. (2017) put it "publication is the currency of academic science" and research that generates positive, novel and clean results is most likely to be rewarded with a publication – incentivizing researchers to seek such results and often biasing their interpretations as a result (Munafò et al., 2017; Nosek et al., 2012). For example, Simmons, Nelson and Simonsohn (2011) demonstrated how easy it would be for a researcher to produce statistically significant evidence for a false hypothesis by exploring various analytic alternatives once the data has been collected. As well, Carp (2012) in reviewing 241 fMRI articles, showed that there are "nearly as many unique analysis pipelines as there were in studies in the sample" when describing the high flexibility of data collection and analysis methods applied by researchers.

Nosek, Spies & Motyl (2012) write that in the current model researchers are less incentivized to support the self-correcting work required to validate the scientific method as they are unlikely to be rewarded for such work which allows false results to "persist in the literature unchallenged, reducing efficiency in knowledge accumulation" (Nosek et al., 2012). False results are costly as they "inspire investment in fruitless research programs and can lead to ineffective policy changes" (Simmons et al., 2011). Begley and Ioannidis (2015) sum up the current state of reproducibility in writing that the "current model of investigator self-regulation and self-censoring does not seem to be serving the scientific community well enough."

Improving Reproducibility

The Center for Open Science, through its Transparency and Openness Promotion (TOP) guidelines, targets the academic reward system by encouraging journals to embrace

reproducibility standards. The TOP guidelines identify the aspects of the research process that should be made openly available to facilitate reproducibility and has been making noticeable progress. As of April 2017, the number of journal and organizations that had become signatories to the TOP guidelines was over 2,900 (https://cos.io/top/) which is a large increase from the 112 reported two years earlier (Nosek et al., 2015).

Recommendations for researchers can be derived from the TOP guidelines and include applying citation norms to data, code, and research materials to recognize them as intellectual contributions; following proposed transparency standards for the reporting of experimental designs, research materials, data sharing, and analytic methods; and pre-registering studies to make all research, particularly the underrepresented null findings, more discoverable and to elicit a clear distinction between confirmatory and exploratory research (Nosek et al., 2015). Clearly distinguishing between "data-independent confirmatory research that is important for testing hypotheses, and data-contingent exploratory research that is important for generating hypotheses" (Munafò et al., 2017) is crucial to reducing the chances of false positive results as "presenting the result of an exploratory analysis as if it arose from a confirmatory test inflates the chance that the result is a false positive" (Button et al., 2013).

Additional articles have provided more domain and problem specific suggestions for researchers seeking to improve the reproducibility of their work. For example, Schmidt (2009) introduces a functional approach to replication for social sciences based on the work of Hendricks (1991) to better determine "what should be changed and what should be kept constant in the design of a replication experiment." Such an approach more explicitly accounts for contextual factors such as participant and researcher characteristics as the competencies and expertise of the researcher can influence the accuracy of the results. Button et al., (2013) in

seeking better powered evidence offers up the recommendations that researchers perform an a priori power calculation to guide their study designs and work collaboratively to increase power and replicate findings. While Simmons et al. (2011) submit six requirements for authors that work to make researcher degrees of freedom more transparent and include authors making an advanced declaration of their rule for terminating data collection, collecting a minimum number of observations, explicitly reporting on all variables and experimental conditions, and reporting both covariate and covariate free results when analyses include covariates. Finally, Stodden et al. (2016) propose Reproducibility Enhancement Principles (REP) for computational research to enable the reproduction of published computational findings on independent systems using the data, code and digital artifacts that have been shared in open trusted repositories as components of the published research. The authors declare that to adequately facilitate reproducibility in computational research "access to the computational steps taken to process data and generate findings is as important as access to data themselves."

The Testing Effect: Effects of Testing on Subsequent Learning

The testing effect has long been identified as an effective learning technique and learning design strategy but it remains under-utilized (McDaniel & Fisher, 1991; Roediger & Karpicke, 2006). A concern that tests are more cognitively demanding than restudying leading to a greater depletion of cognitive resources (Van Merrienboer & Sweller, 2005) available for subsequent learning may contribute to this underutilization.

Testing Effect

Being tested on target material better facilitates retention when compared to the additional study of the same material (Roediger & Karpicke, 2006). The testing effect has been studied in laboratory and educational settings; and has been shown to occur with various types of

learning such as wordlists (Hogan & Kintsch, 1971), general knowledge questions (McDaniel & Fisher, 1991), and geography concepts (Lipko-Speed et al., 2014). Based on the evidence that testing improves long-term retention, Roediger and Karpicke argue that educators should incorporate more tests not for evaluative purposes but rather as learning techniques (Roediger & Karpicke, 2006).

Forward Effect of Testing

The less studied forward effect of testing suggests that the testing of previously studied material can enhance the learning of subsequently presented new material (Pastötter & Bäuml, 2014). Past studies have found that the forward effect of testing can be replicated in the learning of wordlists (Szpunar et al., 2008), narratives (Chan et al., 2009), and videos (Szpunar et al., 2013). In a multiple-list learning study, Szpunar et al. (2008) had participants study five word lists prior to completing a final cumulative recall test. Participants were assigned to either test or non-test groups. The test groups completed a test following the presentation of each single list whereas the non-test groups performed either a distractor task or restudy task. The participants who completed the test activity immediately on lists 1–4 recalled almost twice as many words from list 5 than the non-test groups, thereby demonstrating a forward effect of testing.

Functional search sets

A functional search set "comprises the memory images that have an association with a given set of cues" (Hockley, 2008). Cues provide contextual information to aid in the discrimination of memory images and facilitate retrieval. The cue-overload principle suggests that the "probability of recalling an item declines with the the number of items subsumed by its functional retrieval cue" (Watkins & Watkins, 1976). Szpunar et al. (2008) suggest that "extended study sessions cause a build of proactive interference" which hinders the learning of

subsequently presented new material by overloading the cues associated with the learned material. Proactive interference refers to "previously learned materials hurting our memory for more recently learned materials" (Anderson & Neely, 1996). For example, the functional search set used when attempting to retrieve the material will be larger when prior lists have not been tested thereby reducing the probability of recall. Thus, when no tests have been taken the functional search set includes items from all four previous lists when list 5 is being studied. Whereas when tests have been taken, the functional search set only includes items from list 5 as the lists previously tested on (e.g., lists 1–4) exist in their own functional search sets.

Pastötter et al. (2011) hypothesize that "during list encoding, the memory system binds each list item to the current representation of the subject's internal context" and when retrieval is performed between the study lists, the internal context of the participant is altered – leading to specific context cues being created for each unique list. These list-specific context cues can then be used when future recall is performed thereby enhancing list discrimination and reducing interference between lists.

Encoding proficiency

The build of proactive interference theory is supported by neurocognitive work which reveals that alpha power (8-13 Hz) during item encoding increases as the study material increases within and across lists. This increase in alpha power has been associated with decreasing item encoding proficiency due to memory load and inattention or mind wandering (Pastötter & Bäuml, 2014). Pastötter et al. (2011) conducted a study that used electrophysiological measures of brain activity to investigate the forward testing effect in multiple-list learning. Study results showed that the measures of brain activity in participants who did not perform retrieval activities showed an increase of alpha power from List 1 to List 5

encoding. On the other hand, participants who performed retrieval activities between each list showed no such increase across lists. Furthermore, the changes in alpha-power from List 1 to List 5 encoding predicted subsequent recall performance suggesting that without intermittent retrieval encoding becomes ineffective across lists and with intermittent retrieval the encoding of later lists remains as proficient as the encoding of early lists.

Oscillations, memory and attention

As individual neurons cannot execute complex cognitive operations in isolation, they must form functional networks with other neurons to carry out complex cognitive operations. Neuronal oscillations are thought to reflect fluctuations of neuronal activity that emerge from the synchronous activation of large neuronal ensembles (Roux & Uhlhaas, 2014) and coordinate the activity of distributed neurons during memory operations (Colgin, 2016). Roux and Uhlhaas (2014) in reviewing the relationship between neuronal oscillations and working memory (WM), state that WM can be subdivided into the initial encoding of information, and the maintenance and retrieval of WM items.

The authors suggest that alpha-band activity reflects the active inhibition of task-irrelevant information thereby facilitating encoding of relevant information and gamma-band activity is involved in the maintenance of WM information. In other words, higher alpha may reflect inhibition in task-irrelevant brain regions (Jensen and Mazaheri, 2010). In describing a possible top-down mechanism through which the active inhibition of task-irrelevant information may occur, Jiang, van Gerven and Jensen (2015) have suggested that attention plays a critical role in the gating of information and optimization of cognitive resource application towards memory encoding. The researchers write that “successful long-term memory encoding is reflected by alpha power decreases in the sensory region of the to-be-attended modality and

increases in the sensory region of the to-be-ignored modality to suppress distraction during rehearsal period."

Replication study research questions

Pastötter et al., 2011 investigated the brain activity of the forward effect of testing in a laboratory setting using non-educational material in the form of wordlists, costly equipment and proprietary software. We investigate the forward effect of testing in a classroom setting using non-educational material in the form of wordlists, low-cost equipment and open-source software. Our research questions is: Can the Pastötter et al. (2011) multimodal study of the forward effect of testing be replicated with a low-cost and open-source setup?

Data and Methods

The experiment is a direct replication of a wordlist experiment by Pastötter et al. (2011) that attempts to validate the experimental setup, materials and methods as the contextual variables such as language, profiles of students, equipment used, and data processing techniques applied (Schmidt, 2009) differ from the original study. Due to the battery limitations of the low-cost device, the replication attempt includes the presentation of three rather than five wordlists to be learned by participants. Further, two separate rounds of trials have been run in an attempt to improve the power and reproducibility of the study.

Methods

Our replication experiment had only two conditions: a restudy and retrieval group; three phases: a learning phase, distractor phase, and testing phase; and three learning trials within the learning phases: wordlist trials L1, L2, and L3. Target material was presented in each learning

trial and followed by either a restudy or retrieval activity. Participants from both conditions performed the retrieval activity following the third learning trial (L3) and a final cumulative recall test.

Participants

Fifty-seven Spanish-speaking participants took part in the wordlist experiment. The first round had 22 participants and the second round had 35 participants. However, only 11 of the participants from the first round of trials consented to releasing their data as Open Data and were included in the final study. Thus, the final study includes 46 participants (11 female and 35 male, age: $M = 23.76$, $SD = 6.52$). All participants were placed into a draw for a gift card worth a hundred Euros in exchange for their participation. Written informed consent was received from all participants.

Wordlists, Wordlist Activities and Tests

The target material consisted of three lists of 20 Spanish words (Appendix A) that were drawn from the free EsPal online repository (Duchon et al., 2013). The words were shown one-by-one during the presentation of the wordlists. Prior to each word, a cross was displayed for 1.2s (i.e., the prestimulus period), after which the word was displayed for 2s (i.e., the stimulus onset period). After all words in the wordlist had been presented, participants were prompted to complete an activity. The The wordlist restudy activity prompted participants to restudy a screen displaying all 20 words from the preceding learning trial for 60s. The retrieval activity asked participants to freely recall as many words as they could within 60s and type them on the screen. The retrieval group was not shown whether their answers were correct or not. The L3 quiz and final test were also free recall activities of the same format except the final test asked participants to recall words from all learning trials (L1–L3) and lasted 90s.

Distractor Task

Between the third trial activity and final test, participants completed a 90s-distractor task to limit the availability of sub-vocal rehearsal opportunities – a technique employed to maintain information in working memory (Camos & Portrat, 2015). The distractor task was a 3-back counting task.

Electrophysiological Data

The electroencephalogram signals (EEG) were measured with an Emotiv EPOC® device which is an inexpensive EEG-based, non-invasive Brain-Computer Interface (BCI). The headset consists of a wireless amplifier and 16 wet saline electrodes which include 14 EEG channels and 2 reference electrodes. The electrodes are located and labeled according to the international 10-20 system and the locations of the electrodes are: AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8 and AF4. For the experiment, all the available electrodes of the Emotiv EPOC headset were used. The generated EEG data was digitized using the embedded 16-bit ADC with 128 Hz sampling frequency per channel.

Data Analysis

For the behavioural analysis, the performance on the final segment quiz (L3) between conditions was compared to determine if a relationship existed between prior activity type (restudy or retrieval) and the proficiency of post-activity learning. Statistical analysis was performed using R Studio (RStudio Team, 2015).

Filtering and processing the EEG signal

The EEG signal was acquired and filtered using OpenViBE, an open-source application for using and testing BCIs (Renard, 2010). Firstly, an OpenViBE scenario was created to acquire the EEG signal from the device. The signal was then processed offline through another scenario

which separated the signal into 14 channels, applied a Butterworth 8-13 Hz bandpass filter, extracted epochs, applied signal averaging and then wrote the data to a single file for each participant (Ramirez et al., 2015). The data was then visually inspected to identify bad channels and artefacts which were then removed manually.

Analyzing the EEG data

To determine whether there was a change in alpha power between L1 and L3 encoding, alpha power was calculated by determining the percentage increase or decrease in alpha from the prestimulus period to the stimulus onset period for each word. Next, as "a particular item maintained in working memory will be coded in a highly distributed manner" (Fuster and Bressler, 2012) and higher alpha power may reflect inhibition in task-irrelevant brain regions (Jensen and Mazaheri, 2010), steps were taken to identify the electrode sites that represented the alpha band activity related to the learning of the lists. Initially, alpha power was calculated per participant per word per electrode site. The corresponding EEG data to the first ten artefact-free words from L1 and the last ten artefact-free words from L1 were compared within each participant's trial to determine which electrode sites displayed a significant difference between early trial and late trial encoding activity as "alpha power during item encoding increases with increasing study material, both within and across lists" (Pastötter & Bäuml, 2014). Paired sample Sign Tests were conducted, and only the electrodes (P8, O2) that had displayed a significant increase in alpha power were used in the final analysis, in which alpha power was averaged across the significant electrodes to recalculate the alpha power per word in L1 and L3. An alpha level of .1 was used for the determination of significant electrodes. The percentage power change from L1 to L3 was compared between conditions with the within-subjects factor of list (L1 to L3) and the between-subjects factor of interpolated activity (restudy, retrieval).

Procedure

Participants were told that three learning trials with distinct target material would take place with an activity following each trial. Participants were not aware of which activity they would perform until the instructions for the activity appeared on the laptop screen following each learning trial. An additional activity was to follow the last learning activity and would be followed by a final test. The additional activity was the 3-back counting distractor task. Participants were instructed to study all target material as the final test was cumulative. PsychoPy (Peirce, 2009), an open-source application, was used to present the stimuli to participants, collect behavioural data, and auto transition the experiment based on timed intervals following the initiation of the baseline countdown. EEG signals of participants were recorded from the baseline to the end of the L3 activity. However, the EEG analysis focused solely on the presentation of the wordlists during the learning phase. A 30s backward counting activity preceded each learning trial activity and is depicted in Figure 1 as a distractor task. In total, the wordlist experiment took 25 minutes to run. The overall session including preparation of the EEG setup took about 45 minutes.

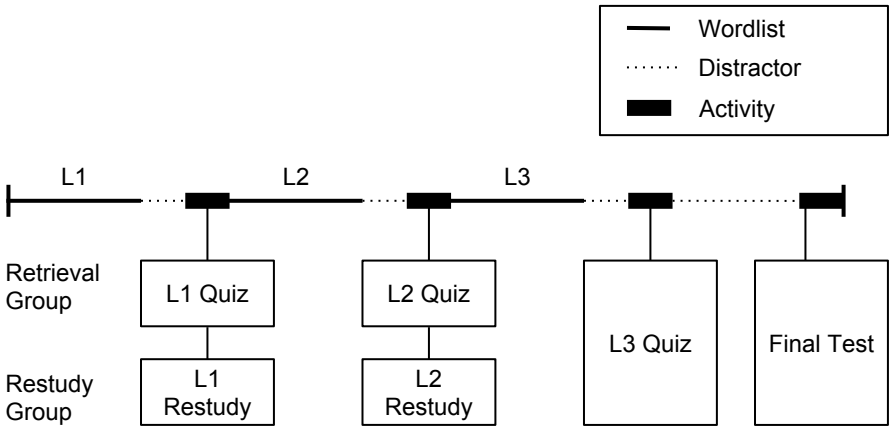

Figure 1. Schematic of the experimental procedure used

Results

Behavioural Results

Immediate recall. Forty-six participants completed the L3 recall test after either completing restudy activities (n = 24) or retrieval activities (n = 22) after L1 and L2. Participants completing the retrieval activities recalled more words on the L3 Test (M = 6.32, SD = 1.84) than the restudy group (M = 2.33, SD = 1.40; U = 509, $p < .001$, 95% CI 3, 5]], d = 2.61). An alpha level of .05 has been used for all statistical tests in the analysis of results.

Final recall. Participants completing the retrieval activities recalled more words on the Final Test (M = 11.91, SD = 3.75) than the restudy group (M = 9.08, SD = 3.35), $t(44) = 2.7$, p = 0.0098, 95% CI [0.72, 4.94], d = 0.8. Table 1 shows the L3 quiz and final test performances per condition.

Physiological Results

Alpha power. Data from thirteen participants could not be used for the physiological data analysis due to a connection failure between the EEG device and the acquisition server which resulted in partial or complete loss of signals recorded. Thus, the analysis of physiological data collected includes 17 participants from the restudy group and 16 participants from the retrieval group. When comparing the changes in alpha power across wordlists L1-L3, the values for the retrieval group were not significantly different from those of the restudy group ($M_{retrieval}$ = 2.82, SD = 14.6; $M_{restudy}$ = -3.89, SD = 7.56; U = 155, p = 0.5043, 95% CI [-1.32, 5.96]), d = 0.24). When making the same comparison with outliers removed, the values for the retrieval group were still not significantly different from those of the restudy group ($M_{retrieval}$ = -1.99, SD = 1.86; $M_{restudy}$ = -2.14, SD = 2.61; U = 80, p = 0.8374, 95% CI [-1.71, 1.94], d = 0.09). A

Spearman's rank-order correlation was run to determine the relationship between L3 Quiz results and the change in alpha power across trials using the data with outliers removed. There was a no statistically significant correlation between the change in alpha power across trials and L3 Quiz results ($r_s(26) = .13$ $p = .512$).

Discussion

We chose to study the testing effect as it offers the potential to serve as a low-barrier method through which immediate and long-term effectiveness of student learning could be improved and has been identified as an effective yet under-utilized learning technique and learning design strategy (McDaniel & Fisher, 1991; Roediger & Karpicke, 2006). We also chose to focus on the forward effect of testing as it potentially addresses a concern that may be affecting the wider adoption of tests as learning tools rather than as assessment tools—a concern that tests are more cognitively demanding than restudying and may deplete cognitive resources negatively affecting subsequent learning. The research conducted involved the collection and analysis of physiological and behavioural data to broadly support the pedagogical integration of formative assessments. Results of past studies (Szpunar et al., 2008; Chan et al., 2009; Pastötter et al., 2011; Szpunar et al., 2013; Pastötter & Bäuml, 2014) and the behavioural results of our direct replication wordlist study show that testing benefits subsequent learning. However, we were not able to reproduce the physiological results of the original study (Pastötter et al., 2011) in which alpha power dynamics differed between the two conditions and correlated with performances on the third learning trial quizzes and final tests. Evaluating whether the replication study shows a statistically significant effect ($P < 0.05$) in the same direction as the original study is a method used by Open Science Collaboration (2015) to evaluate replications. Our inability to reproduce the physiological results, rather than being interpreted as findings

against the work of Pastötter et al. (2011), should be interpreted as a failed attempt to validate our specific multimodal experimental setup. The combination of equipment used, data processing techniques applied, and experimenter expertise in collecting, processing and evaluating EEG data likely impacted the accuracy of the results. Such influencing factors highlight the challenges of both conducting and reproducing robust multimodal work as equipment and levels of expertise often differ and the computational steps taken to process data and generate the findings are not often reported in enough detail to facilitate an exact reproduction.

Reproducibility Challenges

In critically assessing the first round of trials run through a lens of reproducibility to determine whether our work would have adequately supported an independent replication, we identified several issues and through the second round of trials were able to partially address some of the issues. To guide our assessment, reproducibility recommendations that relate to multimodal replication studies (e.g. confirmatory research) were compiled (Table 2).

Study objectives and Sample size

The objective of the study was clearly identified as the direct replication was undertaken to validate an experimental setup for future conceptual replications and to generalize results to a different population. The initial sample size was small with 22 participants in the first replication study attempt as there was a limited time window to conduct the experiment and limited funds available to compensate participants. A second round of trials was run to increase the power of the study. However, the inability to retroactively gain consent and deficiencies with the use of the low-cost device resulted in much of the data being unusable for the physiological analysis.

Methods and Procedures

As a replication study was conducted, the study methods, outcomes and analysis were largely predefined by the original study thereby narrowing the researcher degrees of freedom. Nevertheless, pre-registration was completed for the second round of trials to further restrict researcher flexibility and is accessible online (http://aspredicted.org/blind.php?x=3e8jp9). Pilot experiments were run to become familiar with the experimental setup and device being used. Yet, the EEG data quality was likely affected by a lack of experimenter familiarity in conducting EEG experiments with the Emotiv EPOC and in processing EEG data with open source software. Webb et al. (2015) write that collecting, processing and evaluating EEG data is complex and of critical importance is the signal to noise ratio. Several factors can influence the quality of the EEG signal recording such as motor movements, eye blinks (Bell & Cuevas, 2012), environmental electrical fields, differences in age, gender, phenotype (Cowley et al, 2015), and participant familiarity in performing while being recorded (Webb et al, 2015). In terms of the equipment, the low-cost device was not flexible enough to properly fit all head shapes which affected the quality of the signal collected. Additionally, the Emotiv EPOC device's performance degraded after 25 minutes due to its limited battery life which influenced the designs of the experiment with participants studying three wordlists opposed to the five wordlists in the original study.

Analysis and Reporting

Cumming (2014) states that it is recommended to avoid using null-hypothesis significance testing (NHST) and instead focus on giving "meaningful interpretation of the ES estimates and CIs that give the best answers to your research questions" to best contribute to the building of a cumulative quantitative discipline. The results have been presented in a manner that facilitates power calculations and meta analyses (Lakens, 2013). All experimental conditions, variables and

measures related to the confirmatory analyses have been documented and the final work strives to be Open Science compliant with the data from the study (Beardsley et al., 2017) and the software implementation of the experiment (Beardsley, 2017) publically available online.

Conclusion

When writing about the use of EEG, Webb et al. (2015) argued that to progress from "innovation to significance, protocols and publications should include clear, well justified measurement parameters allowing both for the collection of meaningful, interpretable data, as well as for evaluation and replication by the scientific community." The authors' statement could also be applied to the collection and interpretation of multimodal data in general. As to best make use of the increasingly accessible multimodal data on learning—especially in interdisciplinary investigations—researchers need not only to become familiar with best practices for the modality being investigated and medium being used to conduct the investigation but also need to develop competencies and acquire the tacit knowledge required to reliably collect, process, analyze and report the data. Furthermore, with the procedures needed to produce reproducible research becoming more clearly defined and with a greater number of journals committing to TOP Guidelines, researchers looking to innovatively incorporate multimodal data into their studies should strive to do so sustainably—in a manner that aligns with reproducibility requirements.

References

Anderson, M. C., & Neely, J. H. (1996). Interference and inhibition in memory retrieval. *Memory*, 22, 586.

Baddeley, A. (2012). Working memory: theories, models, and controversies. *Annual review of psychology, 63*, 1-29.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature, 533*(7604), 452-454.

Beardsley, M., Vujovic, M., & Sayis, B. (2017). Open Data For a Multiple Wordlist Learning Task (Version v1.0). Zenodo. http://doi.org/10.5281/zenodo.1123478

Beardsley, M. (2017). PsychoPy implementation of a multiple-list learning task (Version v1.0). Zenodo. http://doi.org/10.5281/zenodo.1120260

Begley, C. G., & Ioannidis, J. P. (2015). Reproducibility in science. *Circulation research*, *116*(1), 116-126. Chicago

Bell, M. A., & Cuevas, K. (2012). Using EEG to study cognitive development: Issues and practices. *Journal of Cognition and Development*, *13*(3), 281-294.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology*, *64*, 417-444.

Bruner, J. S. (1996). *The culture of education*. Harvard University Press.

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

Camos, V., & Portrat, S. (2015). The impact of cognitive load on delayed recall. *Psychonomic bulletin & review, 22*(4), 1029-1034.

Carp, J. (2012). The secret lives of experiments: methods reporting in the fMRI literature. *Neuroimage, 63*(1), 289-300.

Chan, J. C., Thomas, A. K., & Bulevich, J. B. (2009). Recalling a Witnessed Event Increases Eyewitness Suggestibility The Reversed Testing Effect. *Psychological Science*, *20*(1), 66-73.

Christensen, G., Steinmetz, A., Alcorn, B., Bennett, A., Woods, D., & Emanuel, E. J. (2013). The MOOC phenomenon: who takes massive open online courses and why? Available at SSRN 2350964.

Colgin, L. L. (2016). Rhythms of the hippocampal network. *Nature Reviews Neuroscience*.

Cowley, B., Filetti, M., Lukander, K., Torniainen, J., Henelius, A., Ahonen, L., Barral, O., Kosunen, I., Valtonen, T., Huotilainen, M. & Ravaja, N. (2016). The Psychophysiology Primer: A Guide to Methods and a Broad Review with a Focus on Human–Computer Interaction. *Foundations and Trends® Human–Computer Interaction*, *9*(3-4), 151-308.

Cumming, G. (2014). The new statistics: Why and how. *Psychological science*, *25*(1), 7-29.

Duchon, A., Perea, M., Sebastián-Gallés, N., Martí, A., & Carreiras, M. (2013). EsPal: One-stop shopping for Spanish word properties. *Behavior research methods*, *45*(4), 1246-1258.

Fuster, J. M., & Bressler, S. L. (2012). Cognit activation: a mechanism enabling temporal integration in working memory. *Trends in cognitive sciences*, *16*(4), 207-218.

Guo, P. J., Kim, J., & Rubin, R. (2014, March). How video production affects student engagement: An empirical study of mooc videos. In *Proceedings of the first ACM conference on Learning@ scale conference* (pp. 41-50). ACM.

Hendrick, C. (1990). Replications, strict replications, and conceptual replications: are they important? *Journal of Social Behavior and Personality*, *5*(4), 41.

Hockley, W. E. (2008). Memory search: A matter of time. *Learning and memory: A comprehensive reference*, 2, 417-444.

Hogan, R. M., & Kintsch, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning and Verbal Behavior*, *10*(5), 562-567.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLos med*, *2*(8), e124.

Jensen, O., & Mazaheri, A. (2010). Shaping functional architecture by oscillatory alpha activity: gating by inhibition. *Frontiers in human neuroscience*, *4*.

Jiang, H., van Gerven, M. A., & Jensen, O. (2015). Modality-specific alpha modulations facilitate long-term memory encoding in the presence of distracters. *Journal of cognitive neuroscience*.

Klimesch, W. (2012). Alpha-band oscillations, attention, and controlled access to stored information. *Trends in cognitive sciences*, 16(12), 606-617.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in psychology*, *4*.

Lipko-Speed, A., Dunlosky, J., & Rawson, K. A. (2014). Does testing with feedback help grade-school children learn key concepts in science? *Journal of Applied Research in Memory and Cognition*, *3*(3), 171-176.

McDaniel, M. A., & Fisher, R. P. (1991). Tests and test feedback as learning sources. *Contemporary Educational Psychology, 16*(2), 192-201.

Munafò, M.R., Nosek, B.A., Bishop, D.V., Button, K.S., Chambers, C.D., du Sert, N.P., Simonsohn, U., Wagenmakers, E.J., Ware, J.J. and Ioannidis, J.P. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1*, 0021.

Nosek, B.A., Alter, G., Banks, G.C., Borsboom, D., Bowman, S.D., Breckler, S.J., Buck, S., Chambers, C.D., Chin, G., Christensen, G. & Contestabile, M. (2015). Promoting an open research culture. *Science, 348*(6242), 1422-1425. Chicago

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.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716.

Pastötter, B., & Bäuml, K. H. T. (2014). Retrieval practice enhances new learning: the forward effect of testing. *Frontiers in psychology, 5*.

Pastötter, B., Schicker, S., Niedernhuber, J., & Bäuml, K. H. T. (2011). Retrieval during learning facilitates subsequent memory encoding. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(2), 287.

Peirce, J. W. (2009). Generating stimuli for neuroscience using PsychoPy.

Ramirez, R., Palencia-Lefler, M., Giraldo, S., & Vamvakousis, Z. (2015). Musical neurofeedback for treating depression in elderly people. *Frontiers in neuroscience, 9*.

Renard, Y., Lotte, F., Gibert, G., Congedo, M., Maby, E., Delannoy, V., Bertrand, O. and Lécuyer, A. (2010). Openvibe: An open-source software platform to design, test, and use brain–computer interfaces in real and virtual environments. *Presence, 19*(1), 35- 53.

Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*(3), 181-210.

Roux, F., & Uhlhaas, P. J. (2014). Working memory and neural oscillations: alpha– gamma versus theta–gamma codes for distinct WM information? *Trends in cognitive sciences, 18*(1), 16-25.

RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

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.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science, 22(*11), 1359-1366.

Szpunar, K. K., Khan, N. Y., & Schacter, D. L. (2013). Interpolated memory tests reduce mind wandering and improve learning of online lectures. *Proceedings of the National Academy of Sciences, 110*(16), 6313-6317.

Szpunar, K. K., McDermott, K. B., & Roediger III, H. L. (2008). Testing during study insulates against the buildup of proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(6), 1392.

Sternberg, R. J., & Zhang, L. F. (Eds.). (2014). *Perspectives on thinking, learning, and cognitive styles*. Routledge, 232.

Stodden, V., McNutt, M., Bailey, D.H., Deelman, E., Gil, Y., Hanson, B., Heroux, M.A., Ioannidis, J.P. & Taufer, M. (2016). Enhancing reproducibility for computational methods. *Science, 354*(6317), 1240-1241.

Tulving, E., & Watkins, M. J. (1974). On negative transfer: Effects of testing one list on the recall of another. *Journal of Verbal Learning and Verbal Behavior*, *13*(2), 181-193.

Van Merrienboer, J. J., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational psychology review, 17*(2), 147-177.

Webb, S.J., Bernier, R., Henderson, H.A., Johnson, M.H., Jones, E.J., Lerner, M.D., McPartland, J.C., Nelson, C.A., Rojas, D.C., Townsend, J. & Westerfield, M. (2015). Guidelines and best practices for electrophysiological data collection, analysis and reporting in autism. *Journal of autism and developmental disorders*, *45*(2), 425-443.

Wilsch, A., Henry, M. J., Herrmann, B., Maess, B., & Obleser, J. (2014). Alpha oscillatory dynamics index temporal expectation benefits in working memory. *Cerebral Cortex*, 25(7), 1938-1946.

Supporting Information

List of Tables

Table 1. Wordlist Experiment Results 33

Table 2. Multimodal Replication Study Reproducibility Checklist 34

Appendix

Appendix A. Wordlists and Wordlist Ratings 35

Table 1. Wordlist Experiment Results

Condition	*n*	L3 Quiz Words Recalled	Final Test Words Recalled
Restudy	24	2.33 (*SD* = 1.40)	9.08 (*SD* = 3.35)
Retrieval	22	6.32 (*SD* = 1.84)	11.91 (*SD* = 3.75)

Table 2. Multimodal Replication Study Reproducibility Checklist

Topic	Checklist Item	Description	Resources
Study objective(s)	Direct and/or conceptual replication	Direct replication to control for sampling error, artifacts or fraud; or to generalize results to a different population. Conceptual replication to verify the underlying hypothesis of an earlier experiment (Schmidt, 2009).	Theoretical explanation of replication objectives (Schmidt, 2009)
Sample	Sample size declaration	Advance declaration of sample size and how it was determined. Standard practice includes an effect size estimate determined from existing literature (Button et al., 2013).	Software for computing statistical power analyses and effect sizes: G*Power 3 (tiny.cc/gpower3)
Methods	Pre-registration of study	Pre-registration of study design, primary outcome(s) and analysis plan to distinguish data-independent confirmatory research for testing hypotheses, and data-contingent exploratory research for generating hypotheses (Munafò et al., 2017).	Pre-registration repositories: Open Science Framework (https://osf.io/); AsPredicted (https://aspredicted.org/)
Procedures	Multimodal best practices	Familiar with best practices for the collection, processing, analysis and reporting of multimodal data being used or access to an experienced successful experimenter (Schmidt, 2009).	A guide to psychophysiological data collection, processing and analysis for novices (Cowley et al., 2015).
Analysis	Meta-analysis compatibility	Research questions are formulated in estimation terms; effect sizes are being used to answer research; and study results include point estimates and confidence intervals for the chosen effect sizes (Cumming, 2014).	Excel files facilitating the estimation approach to statistics: Exploratory Software for Confidence Intervals, or ESCI (www.thenewstatistics.com)
Reporting	Full disclosure	Documentation of all experimental conditions, variables, measures; stopping guidelines (Button et al, 2013; Simmons et al., 2011); explanation of any departures from the pre-registered study including data exclusions, manipulations, null findings, and explicit identification of any exploratory analysis (Cumming, 2014; Munafò et al., 2017);	
Reporting	Open Science compliance	Data, code, digital artifacts, research materials are available in open repositories; DOIs have been registered and cited in the report; consent forms permit open sharing of data; attribution-only licensing are used for digital scholarly objects; and published or unpublished manuscripts are available in public repositories (Nosek et al., 2012; Stodden et al., 2016).	Open repositories with DOI registrations: Open Science Framework (https://osf.io/); Zenodo (https://zenodo.org/); and figshare (https://figshare.com/)

Appendix A

Wordlists and Wordlist Ratings

Sixty Spanish words (Table A1) were drawn from the free EsPal online repository (Duchon et al., 2013). To produce lists that were of similar difficulty, the words were filtered and grouped based on word frequency, number of letters, number of syllables, and ratings for familiarity, imageability, and concreteness (Table A2).

Table A1. Wordlists

Wordlist 1 (L1)	Wordlist 2 (L2)	Wordlist 3 (L3)
cristiano	reparar	avanzar
morirse	negar	inicio
competir	inglesa	siesta
avance	delta	cruzado
oriente	furia	ritual
derrota	vera	talla
volumen	comedia	conjunto
margen	griego	quitarse
girar	respiro	confesar
coja	miseria	pasaje
quite	ascenso	avisar
confirmar	rendirse	asumir
durar	afecto	cortes
iris	enero	impedir
vengarse	conversar	tumor
reto	editor	dieta
curar	trauma	sensores
bondad	rabia	resistir
ausencia	patria	salario
conflicto	nocturno	consuelo

Table A2. Wordlist Ratings

Condition	Appearance Frequency	Average Number of Letters	Average Number of Syllables	Familiarity Rating	Imageability Rating	Concreteness Rating
L1	11.1	6.5	2.6	5.2	3.8	4.2
L2	11.1	6.3	2.6	5.0	4.2	4.5
L3	11.1	6.7	2.7	5.1	4.0	4.3