



# *A, B, or C* Contrasting: The Influence of a Learning Task on Neurophysiological Correlates of Feedback Processing

Leonie Weindorf, Randi Goertz, Peta Baxter, Frank Léoné

## ABSTRACT

Contrasting similar items has been proposed to enhance learning by increasing the specificity of mental representations. Since feedback is an important component of such tasks, the learning effect of contrasting might be partially related to the neurocognitive mechanisms of feedback processing. Previous electroencephalography (EEG) studies have demonstrated that the P300 and the feedback related negativity (FRN) can indicate whether an instance of feedback leads to successful learning. To investigate whether contrasting influences the effectiveness of feedback processing, we manipulated the orthographic similarity between the answer options on a three-choice vocabulary learning task. EEG was recorded, while participants learned 50 pseudo-translations to Italian words over six blocks. The learning outcome was determined via an immediate and a one-week delayed posttest. Results show better performance during the task for words that were presented with dissimilar distractors (shuffled condition), whereas posttest performance was higher for words that had been studied with similar answer options (sorted condition). The parietal P3b was larger for the sorted compared to the shuffled condition. A larger P3a and smaller FRN to negative feedback were associated with error correction. A larger P3a and smaller FRN to positive feedback were correlated with accuracy on the delayed recognition test. Most of these learning effects were only found for the sorted condition. This indicates that the similarity training might elicit improved memory encoding and attention reflected by the P300 amplitude, as well as enhanced utilization of valence feedback, reflected by the FRN amplitude. Taken together, the results suggest that the memory advantage of contrasting might partially be due to enhanced feedback processing. However, due to the limited sample size (as a result of the corona crisis), no definite conclusions can be drawn, and additional research is needed to corroborate these findings.

Cognitive Neuroscience  
Research Master

Master Thesis by Leonie  
Weindorf

**First reader:**  
Dr. Frank Léoné

**Second reader:**  
Dr. Sybrine Bultena

**Onsite supervisors:**  
Peta Baxter, Randi Goertz

Date: August 24, 2020

**Keywords** *Feedback processing, event-related potentials, declarative learning, orthographic similarity*

## INTRODUCTION

Although research has shown that making errors can be very beneficial to learning, the avoidance of errors is a common and encouraged practice in schools and other educational settings (Metcalf, 2017). Receiving corrective feedback is an essential part of learning from mistakes (Hattie, 1999; Hattie & Timperley, 2007) and whether we learn from feedback has been shown to relate to the way we process it (e.g., Ernst & Steinhauser, 2012; Muller-Gass, Duncan, Tavakoli, & Campbell, 2019). A better understanding of the underlying neurocognitive mechanisms is essential in the process of finding new ways to actively and effectively integrate this aspect of learning into teaching techniques.

This project used Electroencephalography (EEG) to assess feedback processing during a vocabulary learning task. Second language (L2) learning is a well-suited context for such an investigation because it can create a relatively realistic setting (Bultena et al., 2017) and at the same time provides room for manipulation. In the current study we took advantage of this by manipulating the orthographic similarity between the answer options of a multiple-choice word-learning task.

### **Contrasting similar items to enhance specificity of mental representations**

It has been proposed that similar stimuli are represented closely together in so-called “cognitive spaces” (Bellmund, Gärdenfors, Moser, & Doeller, 2018). In these cognitive spaces, “spatially specific cells provide a continuous code” that maps items along certain dimensions of characteristics (Bellmund et al., 2018, p.8). Via this code, stimuli can be represented at different scales, which allows for the representation of individual details as well as generalization. In a language context, this is conceptualized as the mental lexicon, where the dimensions along which similar words are organized are orthography (spelling), phonology (speech sounds), and semantics (meaning). Lexical quality describes the specificity with which orthographic forms are represented and connected to associated semantics and phonology (Perfetti, 2007).

Contrasting similar words during the learning process could be a helpful strategy to improve vocabulary learning by building detailed cognitive representations. Fully specified orthographic representations with constant (rather than variable) letters are important for reading skills (Yang, as cited in Perfetti, 2007) and also seem to play a role in the acquisition of new words (Hart, 2006). Similarity between different words (in terms of form or meaning) has been shown to affect lexical quality (Hart & Perfetti, 2008) and to influence the ease with which these words are learned (e.g., Reder, Liu, Keinath, & Popov, 2016). A common mistake in L2 learning is the confusion of words that are similar in orthographic or phonological form, for example “effect” and “affect”, or “quite” and “quiet” (Llach, 2015). To avoid such mistakes, it can be helpful to contrast small differences during the learning process. “Lexical specificity training”, a method during which new L2 words are taught in phonologically similar pairs, can lead to improved phonological awareness in both first and second language learners (Janssen et al., 2015; van de Ven et al., 2019). Contrasting of orthographically similar words has also been shown to be beneficial for learning (Baxter et al.,

2020). Additionally, contrasting has also been observed to be beneficial in contexts outside of language learning. A study during which subjects had to learn information about specific topics (e.g., the solar system) showed that contrasting similar compared to dissimilar (or less competitive) answer options during a multiple-choice task lead to better performance on a subsequent test (Little & Bjork, 2015).

Llompart and Reinisch (2020) propose that a task requiring the contrasting of similar items over several training blocks increases the attention towards the relevant differences. This enhanced attention might then lead learners to incorporate these detailed differences in the newly acquired mental representations. In the current study, we were specifically interested whether contrasting would influence feedback processing during learning. Generating errors that are similar to the correct answer have been shown to lead to increased retention of word pairs, compared to errors that are not related to the target (Huelser & Metcalfe, 2012). Since learning from errors depends on feedback, the effect contrasting similar words has on the detail of mental representations might not only be due to enhanced attention during the action of contrasting, but also during the processing of feedback.

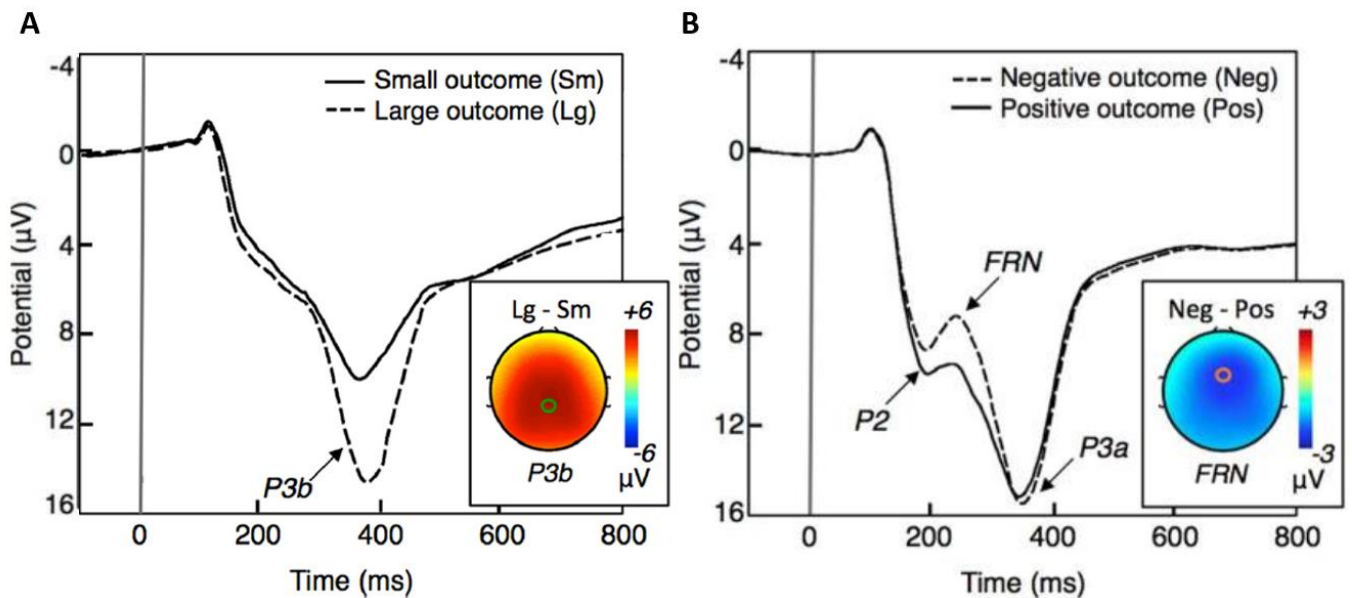
### **ERP components related to feedback processing**

Behaviorally, it is difficult to determine whether contrasting influences the processing of feedback, because feedback processing effects are difficult to distinguish from general task benefits. Event-related potentials (ERPs) have been the method of choice when it comes to investigating feedback-related learning, firstly because they can be measured non-invasively and have a very high temporal precision (Jongsma et al., 2012). Secondly, ERPs can be selectively averaged to compare, for example, items that were later remembered to those that were not remembered, in order to find so-called subsequent memory effects (Sanquist, Rohrbaugh, Syndulko, & Lindsley, 1980). This allows for the assessment of which neural processes are related to learning from feedback and what influences them. Exploring the influence of a similarity manipulation on feedback processing could not only lead to insight about the benefit of contrasting but also provide further evidence regarding the relationship between certain ERP components with learning from feedback.

Mainly two ERP components have been associated with the processing of feedback in the literature: the P300 and the feedback related negativity (FRN). These components will be discussed in the following sections.

#### *The P300*

The P300 component is a positive deflection peaking around 300–600 ms post-stimulus (Figure 1A). This component has previously been researched in many contexts other than feedback processing, and a subsequent memory effect of this component was already found in the 1980s (e.g., Karis, Fabiani, & Donchin, 1984). A larger P300 amplitude is in general thought to reflect successful encoding of stimuli (Polich 2007). This successful encoding is related to both attentional and explicit memory processes (Ernst & Steinhauser, 2012), where the P3a (sometimes also referred to as frontocentral positivity; Themanson et al., 2019, or early



**Figure 1.** “A schematic representation of ERP waveforms typically elicited by outcomes” (taken from San Martín, 2012, p. 2, with a changed order of the two subfigures, creative commons copyright license: <https://creativecommons.org/licenses/by/3.0/>).

frontal positivity; Ernst & Steinhauser, 2012) indicates frontal attention allocation and the parietal P3b reflects memory encoding (Jongsma et al., 2012; Polich, 2007). During feedback presentation these two responses together lead to the subsequent correction of an error (Butterfield & Mangels, 2003) by enabling the evaluation of a response. In a reinforcement learning paradigm this involves “context updating”, where a previous expectation, based on a model of the environment, is evaluated and updated (Ernst & Steinhauser, 2012; Jongsma et al., 2012; Muller-Gass et al., 2019). In a declarative learning context this updating refers to a utilization of the feedback to evaluate the previous stimulus representation or response held in working memory (Ernst & Steinhauser, 2012). The P300 thus reflects the “processing [of] the current stimulus to the degree that it was previously uncertain, index[ing] the cumulative knowledge thereby gained” (Steinemann, Moisello, Ghilardi, & Kelly, 2016, p. 152). The “informative value” of the feedback hence also has an influence (Johnson, 1986; Muller-Gass et al., 2019).

Most studies on feedback processing, including declarative learning studies, report a correlation of the P300 amplitude with successful learning (e.g., Arbel & Wu, 2016; Bultena et al., 2017; Butterfield & Mangels, 2003; Ernst & Steinhauser, 2012; Muller-Gass et al., 2019). For example, in a word-learning study by Ernst and Steinhauser (2012) and in a gender-assignment study by Bultena et al. (2017), a higher first-round P300 amplitude to negative feedback was correlated with a subsequent correction of the error. Further, a visual categorization study by Muller-Gass et al. (2019), who split their participants into successful and less successful learners based on performance accuracy, showed a higher feedback-locked P300 amplitude for more successful learners. The P300 amplitude increased throughout the learning process for the more successful group of learners but not for the less successful one. The P300 amplitude thus seems to reflect individual differences in the effectiveness of feedback processing.

In summary, the P300 can indicate how well feedback is being processed and predict whether this processing will lead to learning. Further, this component is thought to be related to attentional processes, with more attention leading to a higher P3a amplitude (Hillyard, 1985). The P3a and P3b are therefore useful tools in the investigation of attention to detail and feedback processing effectiveness.

#### *The feedback related negativity (FRN)*

The FRN is a negative deflection usually observed at fronto-central electrodes around 200–300 ms after receiving feedback (Figure 1B). Initially, this component was mostly investigated in reinforcement learning studies. Many scholars have found a difference in FRN amplitude to negative compared to positive feedback in this context, usually with a larger FRN to negative feedback (e.g., Holroyd, Hajcak, & Larsen, 2006; Ludowicy, Czernochowski, Weis, Haese, & Lachmann, 2019; Luu, Tucker, Derryberry, Reed, & Poulsen, 2003). This “FRN effect” (difference between positive and negative feedback) has been observed to decrease with learning (Arbel & Wu, 2016; Bellebaum & Daum, 2008; Butterfield & Mangels, 2003; Heldmann et al., 2008; Muller-Gass et al., 2019). Sometimes the FRN has also been found to be larger for positive feedback, which led to doubts that the FRN amplitude solely reflects feedback valence (Themanson et al., 2019).

Several studies have observed higher FRN values for a larger violation of expected outcome (e.g., Chase, Swainson, Durham, Benham, & Cools, 2011; Holroyd & Krigolson, 2007; Themanson et al., 2019; Walsh & Anderson, 2012). In the gender-assignment study by Bultena et al. (2017), a larger difference between the FRN to positive compared to negative feedback was associated with higher certainty ratings, whereas a recent study investigating uncertainty in sensory discrimination found no clear association between FRN amplitude and certainty (Ludowicy et al., 2019). Nevertheless, several scholars suggest that the FRN reflects the difference

between expected and actual outcome (e.g., Ernst & Steinhauser, 2012; Hajcak, Moser, Holroyd, & Simons, 2007; Themanson et al., 2019), as well as an implicit learning process associated with the evaluation of this difference (Bellebaum & Daum, 2008; Chase et al., 2011; Ernst & Steinhauser, 2012).

Whereas several reinforcement learning studies have reported that the FRN amplitude to negative feedback correlates with subsequent correction or change of behavior (e.g., Cohen & Ranganath, 2007; Luu et al., 2003; Philiastides, Biele, Vavatzanidis, Kazzer, & Heekeren, 2010; Van Der Helden, Boksem, & Blom, 2010), such effects are *not* found by most declarative learning studies (Bultena et al., 2017; Butterfield & Mangels, 2003; Muller-Gass et al., 2019). However, several studies by Arbel and colleagues did find learning effects of the FRN. In a study where subjects had to learn paired associations between non-words and novel objects (pictures) during three rounds of a two-choice task, a correlation between the FRN to positive feedback and the learning outcome was observed (Arbel, Murphy, & Donchin, 2014). During a similar task (with five rounds), Arbel and Wu (2016) reported an interaction between the learning curves and slopes of the FRN to negative feedback, namely a steeper learning slope correlated with a steeper decrease in FRN amplitude throughout learning. Further, they found a correlation between successful learning and a large FRN following positive feedback, as well as a small FRN following negative feedback.

Arbel and colleagues proposed a utilization theory, suggesting that the FRN reflects the extent to which feedback is utilized by the learner (Arbel et al., 2014). They also propose that the FRN may account for individual differences in the degree to which relevant information is extracted from feedback (Arbel & Wu, 2016). Related to this utilization theory, Bultena et al.'s (2017) findings suggest differences in the FRN for “more proficient and more perseverant learners” (p.13).

To sum up, the FRN can indicate the detection as well as the expectation of feedback valence (Butterfield & Mangels, 2003), usually with a higher amplitude for negative and unexpected feedback. The FRN also could be related to the utilization of feedback, although this effect has only been found in a very limited number of studies. Since the attention to detail is likely to influence the utilization of feedback, the similarity manipulation in this study could lead to interesting insights in relation to the utilization theory.

## The current study

The current study investigates the influence of contrasting during learning on the effectiveness of feedback processing. For this we applied a multiple-choice vocabulary learning task during which participants learned pseudo-words of which each was orthographically similar to a group of other pseudo-words. Half of the words the participants had to learn were presented with similar answer options (we will call this the sorted condition) and thus required detailed attention for contrasting of small differences. The other half of the words were not presented with the words they were similar to, but instead shuffled in such a way that answer options were always dissimilar (shuffled condition). This design was adapted from a previous related study (Baxter, 2020). To study contrasting in the context of feedback processing, we assessed the differences in feedback-related ERP components between the

two conditions. Since the P300 (and less often the FRN) has been shown to be related to learning, we were particularly interested in differences between the two conditions with regards to correlates of successful learning from feedback. To assess this effectiveness of feedback processing we looked both at whether an instance of negative feedback lead to a subsequent correction of the error on the next trial of that word, and whether it lead to the recollection of the correct translation during the posttests. The aim was to gain a better understanding of the underlying mechanisms of the learning task and the learning benefit of contrasting similar items. Specifically, we investigated the following questions:

1. What is the effect of contrasting on learning (accuracy during the task) and on the learning outcome (accuracy on the immediate and delayed posttests)?
2. Is there an effect of condition on feedback processing, reflected in amplitudes of ERP components (P3a, P3b, and FRN) elicited by negative or positive feedback, and in the change in amplitudes throughout the learning task?
3. Is the amplitude of feedback related ERP components correlated with error correction and learning outcomes and does this effect differ between the two conditions?

Behaviorally we expected to find a higher accuracy during learning and a steeper learning curve for the shuffled condition (with dissimilar answer options), since the task is a little easier and less confusing (H1a). At the same time, however, the contrasting of similar options in the sorted condition should lead to a more detailed encoding. We therefore predicted a higher accuracy in the immediate and delayed posttests for the sorted compared to the shuffled condition (H1b).

Regarding ERP amplitudes, the sorted condition was expected to show a higher P3a amplitude to both negative and positive feedback (H2a) due to the increased attention to detail, as well as a higher P3b amplitude due to more detailed memory encoding (H2b). Additionally, if contrasting leads to more utilization of feedback for learning, we expected higher FRN amplitudes in the sorted condition (H2c).

With regards to a correlation with learning, we expected a correlation with error correction and a subsequent memory effect (i.e., later remembered versus later forgotten) of the P300 in both conditions based on previous research (H3a). We expected these effects to be stronger in the sorted condition, if more detailed attention to feedback is triggered by contrasting and is related to learning (H3b). If this detailed attention also leads to more utilization of feedback, we additionally expected a correlation of the FRN amplitudes with error correction and learning outcomes in the sorted condition (H3c). In the shuffled condition we did not expect a subsequent memory effect of the FRN because no such effects were found in most similar studies (H3d).

## METHODS

### The effect of the corona crisis on this project

For full disclosure, we will briefly report how this project was affected and changed due to the corona pandemic. At the time of lockdown (March 16, 2020), the first subjects were about to be tested. Since on-campus research was suspended, a few changes had to be implemented. Originally, we had

planned to test 30 Dutch native speakers on campus. The full explanation of the original methodology can be viewed in Appendix A and we plan to carry out the original plan once it is possible again. For the sake of finishing this master thesis however, a different approach had to be implemented. A portable TMSi EEG system was used in order to test subjects off campus. This system enabled data collection in line with corona regulations, however these regulations restricted the possible participant pool to two participants. Since these participants were native Italian speakers, the stimuli had to be Italian instead of Dutch. Fifty Italian words were selected based on frequency and word-length, and some of the pseudo-words were changed due to their similarity or equivalence with existing Italian words. Finally, the preprocessing of the EEG data was shifted from BrainVision Analyzer (BrainVision Analyzer, Version 2.2.0; Brain Products GmbH, Gilching, Germany) to FieldTrip (Vers. 2020-0607, Oostenveld et al., 2011; Donders Institute for Brain, Cognition and Behaviour, Radboud University, the Netherlands. See <http://fieldtriptoolbox.org>) in Matlab (Version 2020a; MATLAB, 2020), due to off-campus accessibility. Given the greatly reduced number of participants, results are hence preliminary.

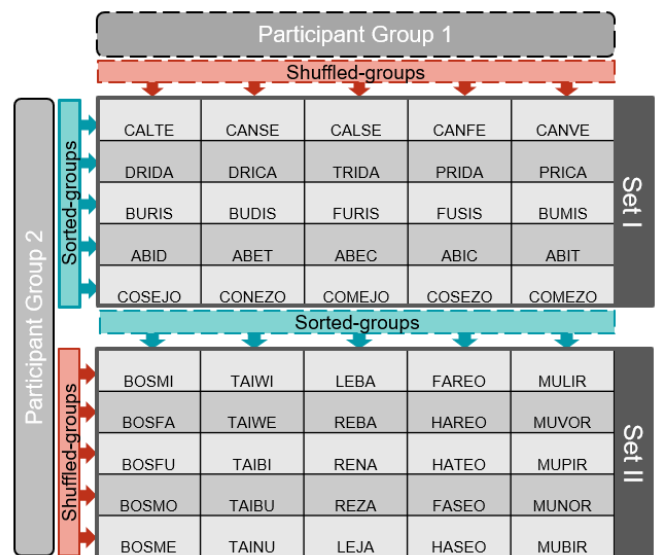
## Participants

Two female subjects between 26 and 27 years of age participated in the study. Both participants had normal or corrected-to-normal vision and reported no colorblindness, psychiatric diagnosis, or previous head injury. The subjects were native speakers of Italian. All subjects gave written informed consent and received 10 € per hour as compensation.

## Stimuli

The stimulus material consisted of 50 Italian words and 50 pseudo-words, with each pseudo-word as a “translation” for one of the Italian words. The Italian words were concrete nouns, four to seven letters long, and their frequency (per million in SUBTLEX-NL; Keuleers, Brysbaert, & New, 2010) was between 0.40 and 70.03 ( $M = 16.03$ ). To guarantee dissimilarity between the Italian words, the Levenshtein distance (indicating the smallest number of single-character changes needed to transform one word into the other; Levenshtein, 1966) between Italian words was kept at a minimum of two. None of the Italian words were semantically similar to each other, based on inspection of their English translations.

The pseudo-words were generated in wuggy.org (Keuleers & Brysbaert, 2010), with Spanish as a base-language. Another native Italian person proofread all of the pseudowords in order to ensure that none of them were real words in Italian or very similar to an Italian word. Nevertheless, all trials with “sale” as a target had to be excluded from the analysis later on, because one of the subjects noticed that it was actually a conjugated form of the Italian verb “salire”. We created ten groups of five orthographically similar words each (e.g., “taiwi”, “taiwe”, “tainu”, “taibu”, “taibi”; see Figure 2). We will call these groups the “sorted-groups”. The Levenshtein distance between words within one sorted-group was between 1 and 2, whereas the Levenshtein distance between words of two different sorted-groups was minimally 3.



**Figure 2.** Illustration of grouping into sorted-groups and shuffled-groups. Half of the participants were in participant group 1 and the other half in participant group 2.

For the control condition, the same 50 words were sorted into a second grouping of ten groups, the “shuffled-groups”. For this reorganization, the ten sorted-groups were first split up into two sets of five groups each, set I and set II. Taking set I, five new groups were generated by selecting one word from each sorted-group. This resulted in five new shuffled-groups, each consisting of five words that were all *not* orthographically similar to each other. The same reorganization was applied to set II. To reduce the influence of nuisance variables, the frequency and length of the Italian words were matched between all sorted-groups. The frequency was also matched between the two sets.

## Design

The similarity manipulation was applied as a counter-balanced, within-subject condition. In one condition, the three pseudo-word answer options were orthographically similar to each other (taken from the same sorted group), called the “sorted condition”, and in the other condition, the pseudo-words were not similar to each other (taken from the same shuffled-group), called the “shuffled condition”. All participants were presented with the same 50 words. However, for half of the participants, the sorted-groups of set I were used, and the shuffled-groups of set II. For the other half of the participants, the shuffled-groups of set I were used and the sorted-groups of set II. For a given participant the words remained in the same condition throughout the whole task.

## Procedure

The subjects participated in two separate sessions. In the first session they performed the vocabulary learning task – during which EEG was recorded – and the immediate posttest. Additionally, participants filled out a general questionnaire and a language experience questionnaire. The delayed posttest took place one week later. No EEG was recorded during this second session. Both the delayed and the immediate posttest consisted of a translation recognition and a translation

production task. All tasks were programmed in PsychoPy3 (Peirce, 2007). The participants sat in front of a 14" HP EliteBook 840 G3 Windows 10 laptop with a 1680x1050 resolution screen and used the laptop keyboard to respond during all tasks.

### Learning task

During the learning task, the participants were asked to learn the pseudo-translation for each of the 50 Italian words. Each of the six blocks consisted of 50 trials, such that all Italian words were tested once per block. To minimize strategies other than contrasting, we instructed the participants to focus on comparing the three answer options and not use mnemonics or any other memorizing techniques. Since the participants were not previously exposed to the pseudo-words, they had to guess in the beginning of the task. Before the first block, participants could familiarize themselves with the task and the type of feedback during six practice trials. To avoid any unwanted effects of the practice trials, these trials did not include any real stimuli, but used placeholders instead, such as "Italian word", and "Option 1".

During each trial (for an example see Figure 3), participants first saw a fixation cross for 500 ms, followed by the Italian probe and the three response options underneath each other. The participants then had 4000 ms to select one of the options via a button press (number key 1, 2, or 3). If they did not respond in time, a message reading "too slow" was displayed, after which the trial continued. After the response, the Italian word remained on the screen as a fixation and to help the subjects remember the Italian word of the current trial. This was followed by a 300 ms blank screen (for the baseline window) and 2500 ms of feedback. The feedback consisted of the correct pseudo-translation, which was presented either in red following incorrect (or too slow) responses, or in green following correct responses. The participants were instructed not to blink during the presentation of feedback. The feedback was followed by a 1000 ms blank screen.

Since there were five words per word-group, this resulted in six possible combinations of answer options per Italian word (the target translation plus six possible pairings of the

remaining four words in the group). Each of these options were presented in one of the six blocks, such that each pseudo-word appeared exactly once as a target and twice as a distractor per block. The order of the trials within each block was pseudo-randomized with the software Mix (Van Casteren & Davis, 2006) with a minimum distance of five trials between trials of the same sorted-group (for words the given participant saw in the sorted condition) or shuffled-group (for words the given participant saw in the shuffled condition). The second Mix-constraint was a maximum repetition of three trials of the same condition. The order of the blocks and the answer options was randomized within the experiment script in PsychoPy.

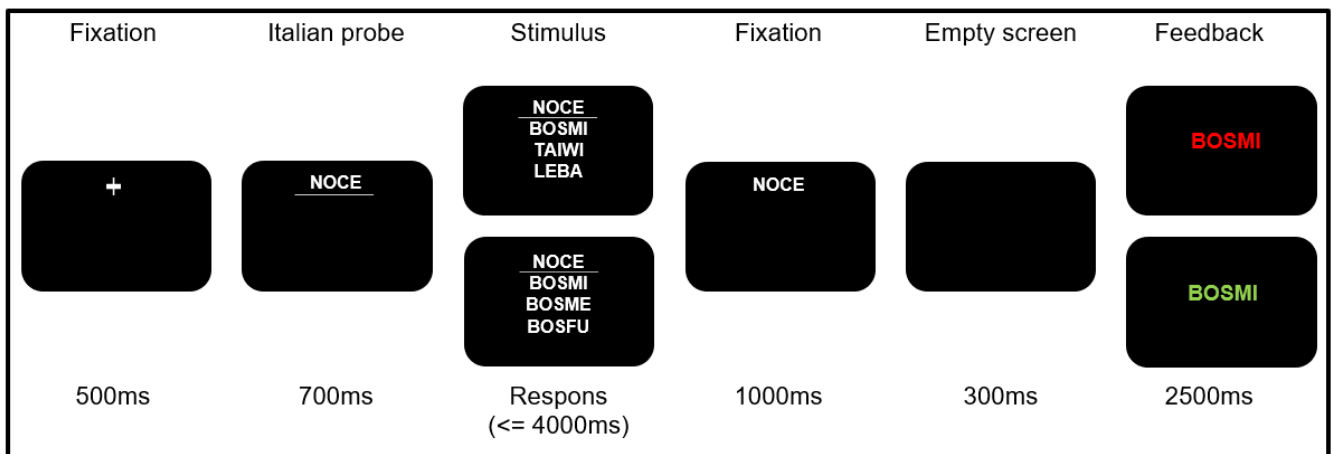
### Posttests

The translation production test consisted of 50 Italian probes presented one-by-one (in a random order). The participants were instructed to type the correct pseudo-translation of the word. They were told to type as much as they remembered and encouraged to guess if they were not sure.

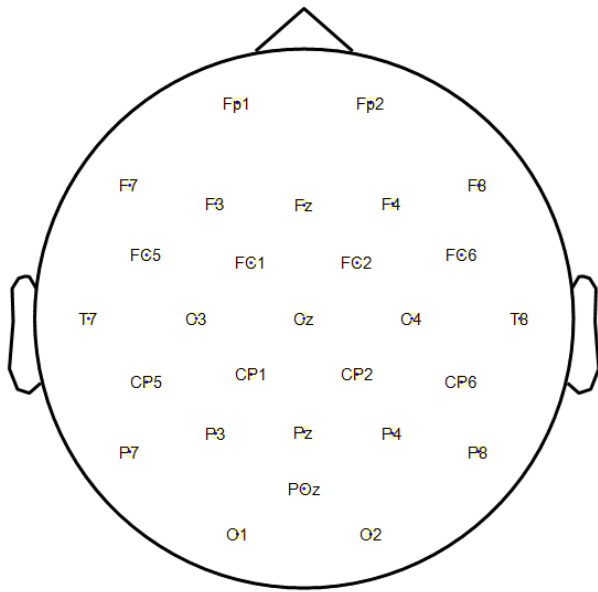
During the translation recognition task, the participants were presented with 50 translation pairs and had to decide whether they were correct or not via a button press (key "n" for no, or key "y" for yes). Nine words from each condition were presented with the correct and 16 words with an incorrect translation. The order of the trials was randomized. The selection of incorrect pairs was pseudo-randomized in PsychoPy with a specific algorithm that ensured counterbalancing of conditions, groups, and similarity of the pairings (see Appendix B for a full explanation of the pair selection). For each condition there was thus an equal number of "similar" (referring to the similarity between the presented – incorrect – translation and the correct translation) and "dissimilar" pairs.

### EEG measurement and preprocessing

The EEG signal was recorded with a TMSi porti system. Twenty-eight electrodes (Fp1, Fp2, F8, F4, Fz, F3, F7, FC5, FC1, FC2, FC6, T8, C4, Cz, C3, T7, CP5, P3, CP1, CP2, P4, CP6, P8, Pz, P7, POz, O1, and O2; Figure 4) were mounted on



**Figure 3.** Trial sequence during the learning task. Stimulus: The upper screen represents a trial in the shuffled condition, and the lower screen is an example of the sorted condition. Feedback: The upper screen is an example of incorrect feedback. In this case the participant selected an incorrect option ("taiwi" or "leba"). The lower screen represents correct feedback, where in this case the participant correctly selected "bosmi" as the translation for "noce" (English translation: "nut"). There was no difference in feedback screens between the two conditions.



**Figure 4.** Illustration of the electrode layout.

an elastic cap according to a 10-20 configuration. Electrodes were online averaged to an additional electrode attached to the wrist with an elastic wristband. The sampling rate was 400 Hz and a high cutoff filter of 108 Hz ( $0.27 \times \text{sampling frequency}$ ) was used. To ensure good signal quality, the overall value of the signal viewer was kept between  $-20$  mV and  $20$  mV during the recording.

The EEG data was preprocessed with the Matlab (Version 2020a; MATLAB, 2020) based FieldTrip software (Version 2020-0607; Oostenveld et al., 2011). The signal was offline re-referenced to an average of all electrodes and a low-pass filter of 30 Hz was applied. The signal was detrended and demeaned. Artifacts were removed via visual inspection of the signal. Feedback-locked epochs were set to 500 ms before and 1500 ms after feedback onset. A time window between 300 ms and 100 ms before feedback onset was used for baseline correction. Trials without responses and practice trials were excluded from the analysis. Single trial ERPs (mean amplitudes) were extracted for statistical analysis in R (Version 4.0.0; R Core Team, 2020). The electrodes for each ERP component were chosen based on previous literature (e.g., Ernst & Steinhauser, 2012; Muller-Gass et al., 2019) and visual inspection of the individual waveforms of all electrodes (similar to e.g., Jongasma et al., 2012). The time windows were chosen based on inspection of the grand averages and individual waveforms of the selected electrodes (e.g., Bultena et al., 2017; Ernst & Steinhauser, 2012). As a result, the FRN was analyzed between 210–270 ms post-feedback at Cz and Fz, the P3a between 365–425 ms post-feedback at Fz and Cz, and the P3b between 460–760 ms post-feedback at Pz.

## Data Analysis

The statistical analysis was performed by means of mixed effects models in R with the lme4 package (Bates et al., 2014; Version 1.1-23). With regards to model fitting and reporting we mainly followed the recommendations by Meteyard and Davies (2020) and Winter (2013). A separate set of models was used to answer each of the three sub-questions outlined in the introduction.

## Model fitting

The fixed and random effects and interaction terms of the initial models were selected based on consideration of study design and factors of interest. All fixed effects were fit as categorical factors, except for learning blocks, which were fit as a continuous variable in the analysis of the behavioral data. The random effects included random intercepts and slopes for item (variation between different stimuli, i.e., words) and participants (variation in performance and ERP signals between participants). Random effects were fitted first, to find a maximal random effects structure that converged (Barr et al., 2013; Meteyard & Davies, 2020). For this we started out with a model including all fixed and random effects of interest, but no interactions. When a model did not converge, random slopes and intercepts explaining zero variance were dropped first. For consistency and comparability, random effects were removed from all models of a subset, if the random effect only converged for some of the models (given the effect explained little variance and did not improve model fit). For models that did not converge with any random effects, regular linear models with only fixed effects were used.

After the model converged, we tested for interactions of interest between fixed effects. Significant two-way interactions were added to the final model if they showed to improve the model fit based on the Akaike Information Criterion (AIC; a reduction by two units was considered as an improvement of fit; Burnham & Anderson, 2004). In case of fixed effects models, the R-squared value was used for model comparison, because no AIC is computed in these models. Significant interactions were followed up on by additionally fitting nested models in both directions, to assess the effects of a factor at each level of the other factor (based on an approach developed by Frömer, Maier, & Rahman, 2018). R Markdown (Allaire et al., 2020) scripts of the selection procedure, including outputs and parameters for all fitted models, can be viewed in the supplementary material.

## Behavioral data

The behavioral data consisted of the accuracy of each response (correct or incorrect response). To address the first research question, three models were fit to the mean proportion correct during the task, the production, and recognition posttests. Due to the binomial accuracy data, the behavioral analysis was conducted with generalized linear mixed models (glmer function), with “logit” link and “bobyqa” optimizer. The p-values and z-values for the behavioral analyses were extracted from the output of the glmer function. For glmer models, the directional interpretation of estimates is not very straightforward (Frömer et al., 2018) and we therefore focused on the data plots when reporting directions of significant effects.

## ERP data

The specific procedure of the statistical analysis of single-trial ERPs was inspired by a method outlined by Frömer et al. (2018). The EEG data was analyzed using the lmer function in R. The p-values for the ERP-models were obtained via Satterthwaite approximations for degrees of freedom, using the lmerTest package in R (Version 3.1-2; Kuznetsova, Brockhoff, & Christensen, 2017). For lmer models the

direction of effects was directly read out from the estimates of the models, as suggested by Frömer et al. (2018).

To compare feedback processing components between the two conditions (question 2), separate models were fit to the amplitude at each electrode of interest per component. This approach was chosen because regional differences was not a main variable of interest, and to avoid overly complex models. Besides an effect of condition, fixed effects of feedback type and block were included because they have been shown to affect signals substantially and meaningfully in similar studies (e.g., Bultena et al., 2017; Ernst & Steinhauser, 2012; Muller-Gass et al., 2019) and should not be collapsed. Additionally, these factors might yield interaction effects with the effect of condition. For the effect of blocks, the blocks were grouped into three different learning phases, namely early (block 1 and 2), middle (block 3 and 4), and late (block 5 and 6). This was done to avoid having too few trials per condition on some blocks. The condition used for the computation of each model was a minimum of seven trials per level (e.g., at least seven positive feedback trials in the sorted condition and early learning phase for each participant).

To assess the correlation between ERP amplitudes and successful learning, we applied two separate approaches. In the first approach we split the negative feedback trials into trials that led to a subsequent correction of the mistake (determined by the response to that same word on the next block) and trials that lead to a repetition of the same mistake. We will call this factor the “subsequent correction effect”. This factor was then fit to the data as a fixed effect. Further, independent of the significance of interactions, we fit a nested model of subsequent accuracy within condition for each component. This nested model approach was implemented uniquely in this case because we were unsure if all subsequent correction effects would be detected when looking at the two conditions combined. This is because the correction of a mistake is not easily comparable between the two conditions since the mistakes differ substantially in size. We were therefore concerned that differences in underlying mechanisms of error correction might obscure the results in a regular model.

The second measure of feedback processing effectiveness was the association of an ERP component amplitude with the accuracy at one of the posttests. For this subsequent memory effect, the trials were split up into those words that were later answered correctly and those answered incorrectly on each of the four posttests separately. The delayed production posttest had to be excluded from this analysis because there were not enough trials (per subject). Further, were there not enough trials of positive feedback on the immediate recognition test. For this posttest we thus only analyzed the negative feedback trials.

## RESULTS

### Behavioral results (Research Question 1)

To assess whether contrasting had an effect on learning we compared the accuracy during the task as well as during the posttests between conditions. The mean proportion correct during each of the learning blocks is illustrated in Figure 5A and was .62 ( $SD = 0.49$ ) in the shuffled condition and .53 ( $SD = 0.50$ ) in the sorted condition. Performance was already above chance level in the first block. This is most likely due to the

fact that within one block a single word appeared twice as a distractor and could therefore be excluded as a target, if it had been the target in a previous trial. A mixed effects model including an interaction of two fixed effects (condition and block) showed an effect of condition and an effect of block, as well as a significant interaction of both (see Table 1). To follow up on this interaction we fitted a model of block nested within condition. This revealed a significant effect of block for the sorted ( $p = .031$ ) and the shuffled condition ( $p < .001$ ).

The mean performance on all posttests was higher in the sorted condition compared to the shuffled condition (see Figure 5B). For the production task, a model including fixed effects of condition and time (immediate or delayed) revealed a significantly lower performance for the shuffled condition as well as a significant effect of time, with a lower performance on the delayed production test.

For the recognition task data, a fixed effects model was used (because all random effects explained zero variance), including an interaction of condition and time. This model revealed a significant effect of condition and a significant effect of time. Although a model with an interaction showed to be a better fit for the data (based on a lower R-squared value), the interaction between condition and time was not significant.

### Interim discussion

In sum, the behavioral data showed that although performance was higher in the shuffled condition during the task (H1a), the sorted condition successfully enhanced learning, as shown in the higher posttest results in both recognition and production tasks (H1b). These findings are in line with our behavioral hypotheses and previous findings (Baxter et al., 2020).

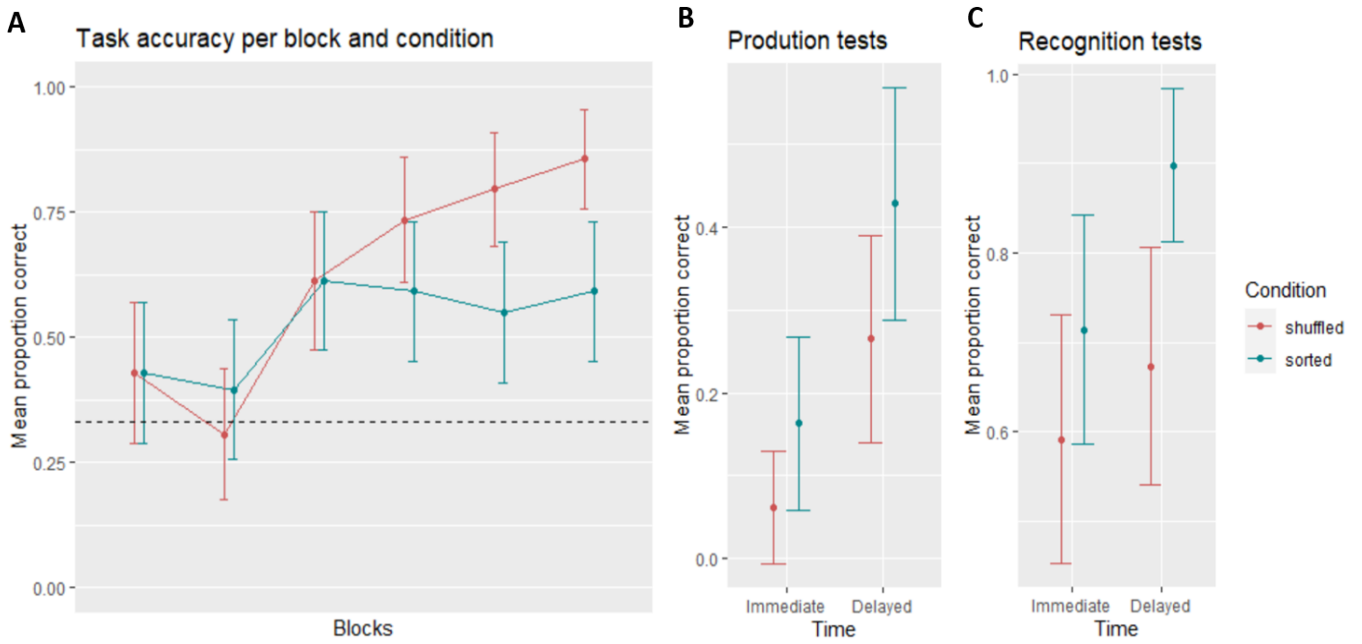
### Electrophysiology during learning (Research Question 2)

Next, we employed ERP-measures (P3a, P3b, & FRN) to determine how the neurocognitive processing of feedback differs with condition, type of feedback, and learning phase. The grand average of the signal at the midline electrodes is illustrated in Figure 6A. Since the amplitude differences are sometimes difficult to see with this many factors, plots of the mean amplitudes grouped by different factors can be viewed in Figure 6B and Appendix C. Linear mixed models with a fixed effect of condition, and an interaction of feedback type (positive, negative) and learning phase (early, middle, late), as well as random intercepts of participants were fit to the data for each component at each electrode (Table 2).

#### P3a amplitude

The P3a amplitude was significantly larger for positive feedback at Cz, but this effect was not found to be significant at Fz (Figure 6B). A significant effect of learning phase was found at both Fz and Cz, with a higher P3a amplitude at later learning phases (Figure 6C). Additionally, an interaction of feedback and learning phase at both electrodes was revealed.

To follow up on these interaction effects, we computed models with nested effects in both directions. Looking at the interaction between learning phase and feedback valence, models of feedback type nested within learning phase revealed a significantly higher P3a amplitude to positive compared to



**Figure 5.** Mean accuracies. Error bars depict 95% confidence intervals. **A:** Mean task accuracy (proportion correct) during each of the six blocks comparing sorted and shuffled conditions. The dashed line indicates chance level performance. **B:** Mean accuracy on the immediate and delayed production test. **C:** Mean accuracy on the immediate and delayed recognition test.

**Table 1.** Effects of condition and block on task accuracy, as well as condition and time on posttest accuracy.

Fixed effects	Task accuracy					Production accuracy					Recognition accuracy				
	<i>b</i>	<i>SE</i>	<i>z</i>	<i>p</i>		<i>b</i>	<i>SE</i>	<i>z</i>	<i>p</i>		<i>b</i>	<i>SE</i>	<i>z</i>	<i>p</i>	
Intercept	-0.42	0.29	-1.45	.147		-0.30	0.33	-0.92	.359		0.90	0.06	14.30	<.001	***
Cond Sh-So	-0.82	0.41	-2.01	.044	*	-0.98	0.41	-2.40	.017	*	-0.22	0.09	-2.53	.012	*
Block	0.16	0.07	2.16	.031	*	-	-	-	-		-	-	-	-	-
Time_De-Im	-	-	-	-		-1.71	0.45	-3.85	<.001	***	-0.18	0.09	0.81	.040	*
Cond*Block	0.38	0.11	3.40	<.001	***	-	-	-	-		-	-	-	-	-
Random Effects	Variance		<i>SD</i>			Variance		<i>SD</i>			Variance		<i>SD</i>		
Item (Intercept)	0.28		0.53			0.96		0.98			-		-		

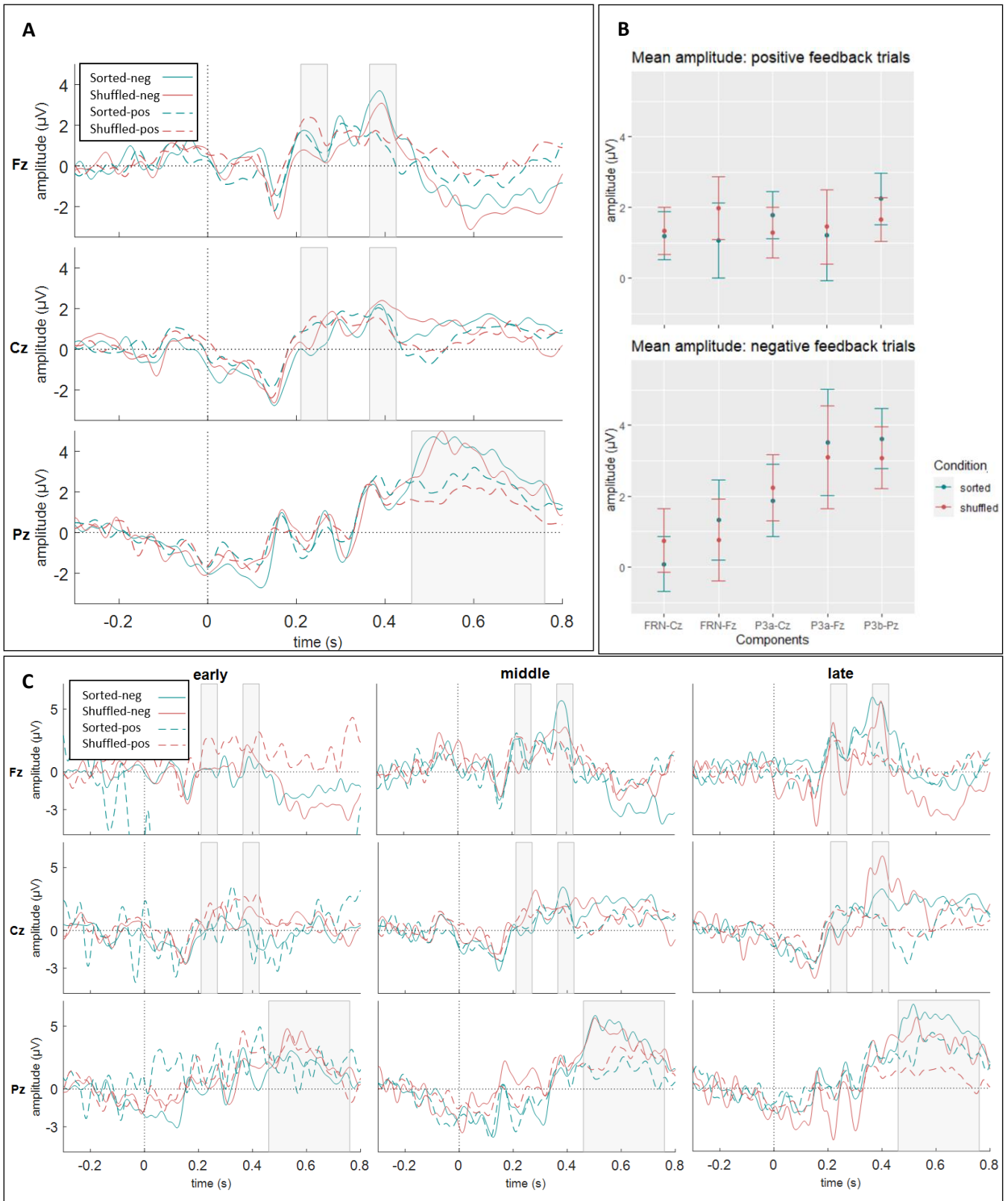
Cond, Condition: Sh = shuffled, So = sorted; Time: De = delayed, Im = immediate  
 \*\*\**p* < .001, \*\**p* < .01, \**p* < .05

negative feedback during the early learning phase at Cz ( $b = 1.55, p = .041$ ), a lower amplitude to positive feedback during the middle learning phase at Fz ( $b = -2.43, p = .027$ ) as well as a lower amplitude to positive feedback during the late phase at both electrodes (Fz:  $b = -4.16, p < .001$ ; Cz:  $b = -2.61, p < .001$ ). A model of learning phase nested within feedback valence showed a significant increase of P3a amplitude from early to middle phase (Fz:  $b = 2.91, p = .011$ ; Cz:  $b = 1.60, p = .030$ ) and from middle to the late phase (Fz:  $b = 4.42, p < .001$ ; Cz:  $b = 2.92, p < .001$ ), for negative feedback. A plot of mean amplitudes grouped by feedback and learning phase showed that for positive feedback the P3a amplitude actually decreased throughout the task (Appendix C, Figure C1A), but this effect did not reach significance.

Furthermore, an interaction of learning phase and condition was found for the P3a amplitude at Cz. This interaction did not improve the model fit enough to be included

in the final model (reduced the AIC by 1.8), but since the interaction was significant, we nevertheless followed up on it. A model of learning phase nested within condition revealed a higher P3a amplitude during later blocks (middle-early:  $b = 2.03, p = .005$ ; late-middle:  $b = 1.55, p = .030$ ) in the sorted condition but not in the shuffled condition. Upon visual inspection of the grand averages, the P3a amplitude during the middle phase seems to be higher for the sorted compared to the shuffled condition (also at Fz), whereas the opposite effect can be seen in the late learning phase. Neither of these effects were significant in a model of condition nested within learning phase, however.

In summary, the P3a was revealed to change from a larger amplitude to positive feedback to a larger amplitude to negative feedback as learning progressed. An increase in P3a amplitude to negative feedback throughout learning was more strongly supported for the sorted condition.



**Figure 6.** Mean ERP amplitudes compared between positive (pos) and negative (neg) feedback as well as between sorted and shuffled conditions. **A:** The grand averages of the EEG signal plotted at each electrode over a time window of 300 ms before and 800 ms after feedback presentation. The grey boxes represent the time windows over which the mean amplitudes were calculated for statistical analysis. **B:** Mean amplitudes of each of the components at the respective electrodes, depicted separately for positive and negative feedback. **C:** Grand averages of the mean EEG signal plotted separately for each block pair and each channel.<sup>a</sup>

<sup>a</sup> In order to be able to spot any differences between conditions, the scale had to be adjusted in such a way the signal for positive feedback in the sorted condition of Fz in the early block pair is not entirely visible (top left panel). This is due to the fact that the signal for this condition showed much more negative values compared to the other conditions and blocks (probably caused by the very noisy signal of one subject on the first two blocks, and the very small amount of trials remaining after cleaning).

**Table 2.** Effects of condition, block, and feedback type on ERP amplitudes

P300	P3a Fz				P3a Cz				P3b Pz						
	<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>	<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>	<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>			
Intercept	0.84	1.90	0.45	.689	-0.15	0.98	-0.02	.988	2.24	0.75	3.01	.018	*		
FB_pos-neg	0.35	1.18	0.30	.766	1.53	0.75	2.03	.043	*	-0.26	0.70	-0.37	.712		
Cond_Sh-So	0.12	0.61	0.19	.846	1.28	0.73	1.74	.083		0.11	0.69	0.16	.873		
BP_mid-ear	2.91	1.14	2.54	.011	*	2.73	0.88	3.12	.002	**	1.74	0.82	2.13	.034	*
BP_lat-mid	4.42	1.20	3.67	>.001	***	3.51	0.87	4.05	<.001	***	3.00	0.81	3.72	<.001	***
Cond*BP(mid)	-	-	-	-		-	-	-	-		0.11	0.93	0.12	.902	
Cond*BP(lat)	-	-	-	-		-	-	-	-		-1.60	0.94	-1.70	.091	
FB*BP(mid)	-2.78	1.61	-1.73	.084		-1.95	1.03	-1.90	.060		-1.57	0.96	-1.63	.104	
FB*BP(lat)	-4.51	1.65	-2.74	.006	**	-4.16	1.07	-3.00	.001	***	-2.50	1.00	-2.50	.013	*
<b>Random Effects</b>	<b>Variance</b>		<b>SD</b>		<b>Variance</b>		<b>SD</b>		<b>Variance</b>		<b>SD</b>				
Participant (Intercept)	5.93		2.43		1.20		1.10		0.50		0.71				
Residual	50.19		7.08		20.63		4.54		17.95		4.24				
<b>FRN</b>	<b>FRN Fz</b>				<b>FRN Cz</b>										
<b>Fixed effects</b>	<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>	<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>							
Intercept	0.11	1.12	0.10	.927	0.05	0.97	0.05	.961							
FB_pos-neg	0.33	0.55	0.60	.548	1.09	0.38	2.85	.005	**						
Cond_Sh-So	0.30	0.52	0.57	.571	0.30	0.36	0.84	.401							
BP_mid-ear	0.78	0.67	1.26	.246	0.09	0.46	0.19	.848							
BP_lat-mid	1.46	0.67	2.16	.031	*	-0.22	0.46	-0.48	.630						
<b>Random Effects</b>	<b>Variance</b>		<b>SD</b>		<b>Variance</b>		<b>SD</b>								
Participant (Intercept)	1.84		1.36		1.84		1.36								
Residual	35.60		6.01		35.60		5.97								

FB, Feedback: pos = positive, neg: negative; Cond, Condition: Sh = shuffled, So = sorted; BP, block pair: mid = middle, ear = early, lat = late  
\*\*\**p* < .001, \*\**p* < .01, \**p* < .05

### P3b amplitude

Neither the condition nor the feedback valence had a significant effect on the P3b amplitude at Pz. We observed a significantly larger P3b amplitude for later learning phases. A significant interaction of feedback type and learning phase, as well as an interaction of condition and learning phase was found.

As a follow-up on the interaction of feedback type and learning phase, a nested model of feedback type within learning phase showed a lower P3b amplitude for positive compared to negative feedback trials during middle ( $b = -1.80$ ,  $p = .006$ ) and late ( $b = -3.11$ ,  $p < .001$ ) learning phases. Further, a significantly higher P3b amplitude on later compared to earlier blocks (middle-early:  $b = 1.79$ ,  $p = .009$ ; late-middle:  $b = 2.47$ ,  $p < .001$ ) was found for negative feedback, but not for positive feedback.

To follow up on the interaction of condition and learning phase, nesting condition within learning phase revealed significantly lower P3b amplitudes ( $b = -1.80$ ,  $p = .005$ ) for the shuffled compared to the sorted condition during the late learning phase. Significantly higher P3b amplitudes on the late

compared to middle learning phase ( $b = 1.86$ ,  $p = .005$ ) were found for the sorted condition only.

To sum up, the P3b to positive feedback was found to be larger in the sorted compared to the shuffled condition, although only later in the learning process. The P3b amplitude was higher to negative feedback compared to positive feedback and the amplitude to negative feedback increased throughout the task.

### FRN amplitude

The FRN was found to be maximal at Fz and Cz electrodes, with a maximal difference between positive and negative feedback at Cz. The FRN amplitude at Cz was significantly larger (i.e., more negative) for negative feedback compared to positive feedback. Note that the FRN is a negative component, and negative estimates hence reflect an increase in FRN amplitude. Neither the effect of condition nor of learning phase were significant at this electrode. The FRN at Fz was found to be significantly larger for the late compared to the middle learning phase. In summary, the FRN was larger to negative feedback and increased throughout learning.

### *Interim discussion*

In this section we assessed the influences of different factors on the amplitudes of feedback related ERP components to investigate whether feedback processing differs between the sorted and shuffled condition. The P3b amplitude was found to be larger in the sorted condition compared to the shuffled condition. This finding is in line with our hypothesis of a higher P300 amplitude in the sorted condition (H2b). However, this effect was only found with regards to positive feedback in the late learning phase and no such effect was found for the earlier P3a component (contrary to H2a). No effect of condition was found with regards to the FRN amplitude (contrary to H2c).

We also assessed the effects of feedback type and learning phase on the ERP components, as well as interactions between these effects and condition. The P3a had a higher amplitude for positive compared to negative feedback in the early learning phase, whereas in the later phases the P3a and P3b were larger on negative feedback trials. Implications of this finding will be addressed in the general discussion. The FRN was larger for negative compared to positive feedback. This is in line with previous literature (e.g., Holroyd et al., 2006; Ludowicy et al., 2019; Luu et al., 2003) and confirms the valence sensitivity of the FRN in our task. Both the P300 and FRN amplitudes to negative feedback increased throughout the task, while for the P3a at Cz and the P3b this increase in amplitude was only found for the sorted condition. This again suggests an effect of condition on the P300 amplitude, with an enhanced P300 in the sorted condition (H2a and H2b).

### **Effectiveness of feedback processing (Research Question 3)**

Next, we assessed the correlation of ERP amplitudes with successful learning to investigate whether this effect differed between conditions. For this, we looked for subsequent correction effects (correlation of amplitudes to negative feedback with accuracy on the next trial of the same word) and subsequent memory effects (correlation of ERP amplitudes with accuracy on the posttests).

#### *Correlation of ERP amplitudes with subsequent correction*

To assess the subsequent correction effect (correlation of amplitudes to negative feedback with accuracy on the next trial of the same word), models with condition and subsequent accuracy as fixed effects and subject random intercepts were fit to the amplitude of each component at the corresponding electrodes (see Figure 7A and Table 3). For the P3a, a significant effect of subsequent accuracy, with a higher P3a amplitude for subsequently corrected trials, was found at Cz.

An interaction between condition and subsequent correction was not significant for the P3a ( $p = .968$ ). A nested model of P3a amplitudes at Cz was not significant for the sorted nor the shuffled condition (sorted trials:  $b = 1.67$ ,  $p = .103$ ; shuffled:  $b = 1.74$ ,  $p = .120$ ). Upon visual inspection of the grand averages the subsequent accuracy effect appears to be present at Fz, but only in the sorted condition. However, a model of subsequent accuracy nested within condition for P3a amplitudes at Fz was not significant for either condition (sorted:  $p = .299$ , shuffled:  $p = .913$ ). No significant effect of subsequent accuracy on P3b amplitudes was found.

A significant effect of subsequent correctness was found for the FRN at Cz, where a lower FRN amplitude was associated with a subsequent correction of an error. The interaction between condition and subsequent accuracy was not significant ( $p = .238$ ), nevertheless in a nested model of subsequent accuracy within condition the effect was only significant for the sorted condition ( $p = .017$ ). At Fz no significant effect was found for the FRN amplitude.

To summarize these findings, a larger P3a and a smaller FRN were associated with the subsequent correction of an error. The FRN (subsequent correction) effect was only significant in the sorted condition.

#### *Subsequent memory effects*

To look for subsequent memory effects (correlation of ERP amplitudes with accuracy on the posttests), models included feedback type, posttest accuracy, and condition as fixed effects, as well as subject random intercepts (see Table 4).

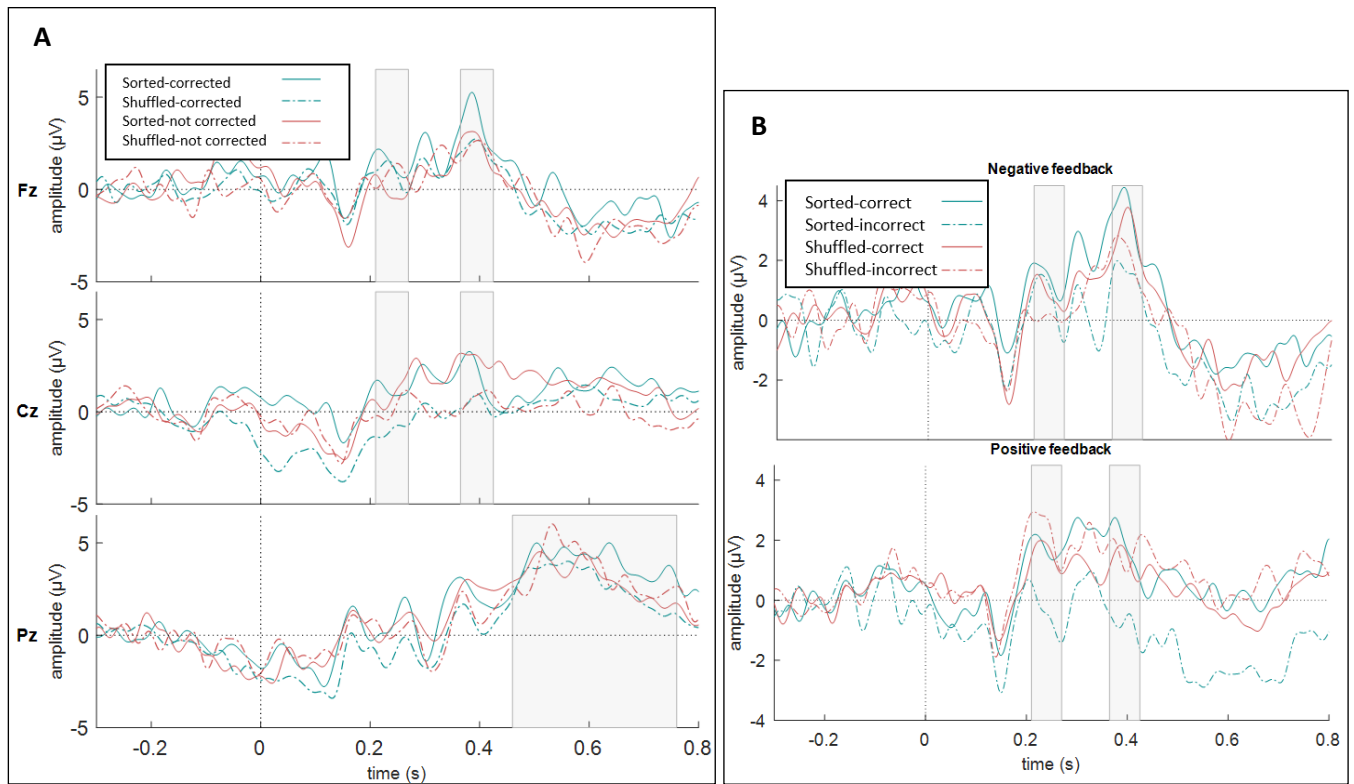
A significant effect of accuracy during the delayed recognition test was found for the P3a amplitude at Fz (see Figure 7B). A higher P3a amplitude was revealed for trials of items that were answered correctly at the delayed recognition task, compared to those answered incorrectly. A three-way interaction was significant, and a follow-up nested model showed a significant effect in the sorted but not the shuffled condition. This effect revealed a higher P3a amplitude to positive feedback for sorted-condition items that were answered correctly at the delayed recognition task compared to sorted-condition items that were answered incorrectly ( $b = 2.72$ ,  $p = .039$ ).

Further, a significant three-way interaction was revealed for the FRN at Fz. In a nested model, a lower FRN amplitude for correct compared to incorrect responses at the posttest ( $b = 2.30$ ,  $p = .036$ ) was found for positive feedback in the sorted, but not in the shuffled condition. Further, this model showed a lower FRN amplitude to positive feedback for the shuffled compared to the sorted condition ( $b = 3.08$ ,  $p = .009$ ). No significant subsequent memory effects were found for the other posttest (for results of the immediate production and immediate recognition models see Appendix D).

In summary, the P3a to positive feedback was revealed to be larger for items correctly recalled at the delayed recognition task. This effect was only found for the sorted condition. A smaller FRN to positive feedback was also associated with correct recall at the delayed recognition task in the sorted condition only.

### *Interim discussion*

This section assessed the correlation of feedback-processing related ERP components with learning in the two conditions. The subsequent correction and subsequent memory effects of a larger P3a were in line with our hypothesis (H3a). In our predictions, we did not specify a direction of effects regarding a correlation of the FRN amplitudes with successful learning. Nevertheless, in light of the utilization theory of the FRN, it is a little surprising that a smaller amplitude rather than a larger amplitude was associated with error correction and subsequent memory (H3c). This will be further discussed in the general discussion. The learning effects with regards to the FRN were only found in the sorted condition and not in the shuffled condition (H3d), which



**Figure 7.** Leaning effects. **A:** Negative feedback trials that were followed by a correction of the error on the next block (corrected) compared to negative feedback trials that were followed by another instance of negative feedback on a trial for the same word on the next block (not corrected). **B:** ERPs at Fz of correct compared to incorrect items at the delayed recognition task. Upper: positive feedback trials, lower: negative feedback trials.

**Table 3.** Effects of subsequent correction on ERP amplitudes

<b>P300</b>	<b>P3a Fz</b>				<b>P3a Cz</b>				<b>P3b Pz</b>				
<b>Fixed effects</b>	<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>	<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>	<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>	
Intercept	2.65	2.30	1.15	.350	0.56	0.85	0.66	.546	3.03	0.67	4.52	.006	**
SubCorr_C-N	0.94	1.11	0.85	.399	1.71	0.75	2.27	.024	* 0.52	0.67	0.78	.439	
Cond_Sh-So	-0.79	1.11	-0.71	.478	0.45	0.75	0.60	.553	-0.14	0.67	-0.21	.834	
<b>Random Effects</b>	<b>Variance</b>			<b>SD</b>	<b>Variance</b>			<b>SD</b>	<b>Variance</b>			<b>SD</b>	
Participant (Intercept)	9.00			3.00	0.74			0.86	0.34			0.59	
Residual	56.92			7.54	26.07			5.11	20.50			4.53	
<b>FRN</b>	<b>FRN Fz</b>				<b>FRN Cz</b>								
<b>Fixed effects</b>	<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>	<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>					
Intercept	1.31	1.19	1.10	.349	-0.67	0.64	-1.05	.343					
SubCorr_C-N	-0.48	0.90	-0.54	.592	1.41	0.63	2.24	.027	*				
Cond_Sh-So	-0.65	0.90	0.72	.473	0.58	0.63	0.91	.364					
<b>Random Effects</b>	<b>Variance</b>			<b>SD</b>	<b>Variance</b>			<b>SD</b>					
Participant (Intercept)	1.84			1.36	0.31			0.56					
Residual	37.60			6.13	18.46			4.30					

SubCorr, Subsequently corrected: C = corrected, NC = not corrected, Cond, Condition: Sh = shuffled, So = sorted  
 \*\*\**p* < .001, \*\**p* < .01, \**p* < .05

**Table 4.** Effects of delayed recognition accuracy on ERP amplitude<sup>b</sup>

<b>P300</b>		<b>P3a Fz</b>				<b>P3a Cz</b>				<b>P3b Pz</b>				
<b>Fixed effects</b>	<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>	<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>	<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>		
Intercept	1.95	1.87	1.05	.383	2.04	0.93	2.19	.110	3.48	0.67	5.19	.004	**	
FB_pos-neg	-1.67	0.64	-2.61	.009	**	-0.35	0.42	-0.84	.401	-1.50	0.39	-3.89	<.001	***
Cond_Sh-So	0.08	0.64	0.13	.895		-0.21	0.41	-0.50	.614	-0.49	0.38	-1.28	.200	
RD_C-I	1.45	0.66	2.19	.029	*	-0.16	0.43	-0.37	.709	0.36	0.40	0.90	.368	
<b>Random Effects</b>	<b>Variance</b>			<b>SD</b>	<b>Variance</b>			<b>SD</b>	<b>Variance</b>			<b>SD</b>		
Participant (Intercept)	5.94			2.44	1.28			1.13	0.51			0.72		
Residual	51.16			7.15	21.58			4.65	18.59			4.31		
<b>FRN</b>		<b>FRN Fz</b>				<b>FRN Cz</b>								
<b>Fixed effects</b>	<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>	<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>						
Intercept	0.27	1.11	0.24	.820	0.26	0.97	0.27	.805						
FB_pos-neg	0.62	0.54	1.16	.245	1.05	0.37	2.86	.004	**					
Cond_Sh-So	0.36	0.53	0.67	.501	0.26	0.37	0.72	.471						
RD_C-I	0.67	0.55	1.21	.225	-0.34	0.38	-0.89	.374						
<b>Random Effects</b>	<b>Variance</b>			<b>SD</b>	<b>Variance</b>			<b>SD</b>						
Participant (Intercept)	1.74			1.32	1.53			1.24						
Residual	35.82			6.10	16.94			4.12						

FB, Feedback: pos = positive, neg: negative; Cond, Condition: Sh = shuffled, So = sorted; RD, delayed recognition: C = correct response, I = incorrect response  
\*\*\*p < .001, \*\*p < .01, \*p < .05

supports our hypothesis of the enhancement of feedback utilization in this condition.

## GENERAL DISCUSSION

The present study investigated feedback processing during a word-learning task that was aimed at manipulating the attention to detail required during learning. For this manipulation, words in the sorted condition were presented with similar answer options, and words in the shuffled condition with dissimilar ones. The main goal was to elucidate whether contrasting would influence feedback related ERP components and to explore whether successful learning during such a contrasting task is related to feedback processing.

Performance during the task and the posttests differed between conditions. During the task, the performance was higher and the learning curve steeper for the shuffled condition. The performance on the immediate, as well as the one-week delayed posttests on the other hand, was higher in the sorted condition. Although the task was thus more difficult with similar answer options, this manipulation lead to an improved memory of the words right after the task as well as long-term.

To assess whether and how this memory benefit of the similarity manipulation might be related to a difference in the processing of external feedback, we assessed the amplitude of relevant ERP components during feedback presentation. In the following sections, we discuss the identified differences in these ERP components between the two conditions in terms of valence and learning phase effects on the amplitudes as well as their correlation with learning. The small sample size of the study substantially limits the validity of our findings, nevertheless in the following we discuss possible explanations and potential implications of our findings.

### Feedback processing

Examination of the EEG signal revealed a P300, which was observed as an earlier fronto-central P3a and a later parietal P3b component. Additionally, a fronto-centrally peaking FRN component was visible around 210–260 ms post-feedback.

#### Condition effects on ERP amplitudes

The only direct comparison yielding a difference of ERP amplitudes between sorted and shuffled condition was a higher P3b in the late learning phase. This finding supports the

<sup>b</sup> Including fixed effects of condition and feedback in the posttest models, automatically lead to repeated analyses of these effects and it has to be noted that some of these results were different to those of the previously reported models (electrophysiological results section). Effects that were different from those previously reported are briefly described here. The P3a was found to be significantly smaller for positive feedback ( $b = -1.51, p = .021$ ) at Fz in the immediate production model. Whereas the P3a at Cz was not significantly modulated by feedback type in the immediate production, nor the delayed recognition model. The P3b at Pz was found to be significantly smaller for positive feedback in both the model with immediate production accuracy ( $b = -1.48, p = .002$ ) and in the model with delayed recognition accuracy ( $b = -1.50, p = <.001$ ).

hypothesis that in the sorted condition, more feedback evaluation and context updating resources might be recruited (H2b). Since the P3b component is thought to be related to memory processing, this could also partially explain the higher performance on the posttest, although in this current study no direct relationship between P3b amplitude and posttest performance was found.

The current data suggests that there might be additional, more complex differences in ERP amplitudes between conditions, which did not reach significance in the small sample. In fact, when looking at the signal of both participants separately, there seem to be substantial, but opposite effects of the two conditions on P3a and FRN amplitudes (see Appendix E). In a larger study, this effect might also on average be zero, or it might be stronger in one direction. These (potential) differences could also be related to performance or some control variable.

### *Feedback valence effects on the P300*

In the early learning phase, the P3a was larger for positive compared to negative feedback, whereas in the latter two phases both the P3a and the P3b had a higher amplitude on negative feedback trials. In previous research the P300 is sometimes found to be larger for positive feedback (e.g., Ernst & Steinhauser, 2012) and sometimes larger for negative feedback (e.g., Bultena et al., 2017). Interestingly, Arbel, Hong, Baker, and Holroyd (2017) also report a larger P3a for positive feedback in the early learning phase and a larger P3a to negative feedback for the middle and late learning phases. This shift might be related to stimulus probability, since several other studies have reported a higher P300 to less probable stimuli (e.g., Arbel & Wu, 2016; Ernst & Steinhauser, 2012; Jongsma et al., 2012). A support for this reasoning can be observed in the behavioral data of the present study, where the occurrence of positive feedback switches from below to above 50%, exactly at the transition between early and middle learning phase (between block 2 and 3).

Possibly, this is also related to a rise in confidence throughout the task. A large P3a amplitude has previously been reported for trials of high “meta-memory mismatch” (meaning “high-confidence errors and low-confidence corrects”; Butterfield & Mangels, 2003, p. 804). Relatively high P300 amplitudes to positive feedback for the early phase could thus be due to a low confidence in the beginning of the task, and relatively high P300 amplitudes to negative feedback could be related to a rising confidence as learning progresses.

Another reason for the switch from a higher P3a amplitude following positive feedback to a higher P3a amplitude following negative feedback, could be an increase in contextual updating throughout the learning task. In the beginning of the task, the words are unknown to the learner, and an instance of negative feedback thus cannot trigger an evaluation of the previous stimulus representation. In fact, not even the response might be remembered well enough to evaluate its content. As the words become more and more familiar to the learners throughout the task, feedback processing might become more focussed on comparing response and feedback, as well as updating the mental representation of the word, rather than solely processing the correctness of the response.

It is important to note that the results with regards to the P300 are not as clear as one would hope, which is at least

partially due to the small number of participants. The effect of feedback valence on P3a amplitude was not found in all posttest models. This is probably due to the changing directionality of the relationship between amplitude and feedback valence throughout the task, which is not taken into account in these posttest models.

### *Learning phase effects and their interactions with condition*

Both the P300 and FRN amplitudes to negative feedback increased, rather than decreased throughout the task. This is similar to a finding by Bultena et al. (2017), who argue that this is the case when feedback continues to be relevant to the learner (see also Heldmann et al., 2008). In terms of the P3a, the increase in amplitude was more strongly supported for the sorted than for the shuffled condition. This could also be related to an increase in attention and updating. As elaborated above, the attention to and evaluation of negative feedback might increase as words become more familiar. This could especially be the case in the sorted condition, because even more attention is needed to compare the small differences between the selected response and the correct answer presented by the informative feedback. An increasing P300 amplitude throughout the learning process has also been associated with better learning performance in general, since such an increase was only observed for more successful learners by Muller-Gass et al. (2019). The increase in P300 in the current task could thus also be related to the learning benefit of the sorted condition.

The FRN is often associated with expectancy (e.g., Bultena et al., 2017; Chase et al., 2011; Holroyd & Krigolson, 2007; Themanson et al., 2019; Walsh & Anderson, 2012). It is possible that the increase in FRN amplitude is due to a decreasing expectancy of negative feedback as performance improved throughout the task.

### **Correlation of feedback processing and learning**

Lastly, we investigated the correlation of ERP amplitudes with successful learning. During this task, a larger P3a and a smaller FRN to negative feedback were found on trials that lead to a correct response in the next block. Further, the P3a to positive feedback was larger, and the FRN smaller for words that were recalled during the delayed recognition task. Both components were thus associated with the correction of errors and the learning from positive feedback.

### *The P300 and successful learning*

Our subsequent correction finding of the P3a confirms that the well-studied association of the P300 amplitude with successful encoding applied to the present task as well. It should be noted that most studies focus on the amplitude during the first round of learning and its correlation with the performance on the next block or the posttest. In the present study this focus on the first round was not possible, due to too few trials per condition per block. Instead we divided trials of all blocks into ones that led to positive versus negative feedback on the next block (subsequent correction; similar to a design by Steinemann et al., 2016) and in a separate analysis into items with a correct versus an incorrect response on one of the posttests (subsequent memory). The effect of subsequent correction for negative feedback is similar to that found in

other declarative and L2-learning studies (e.g., Bultena et al., 2017; Ernst & Steinhauser, 2012). It suggests that more attention to negative feedback leads to an evaluation of the response and a correction of the error.

The finding of a higher P3a amplitude to positive feedback being associated with the performance on a delayed posttest is more novel in a declarative learning context, however. Several studies focus solely on negative feedback when assessing correlation with learning (e.g., Bultena et al., 2017; Butterfield & Mangels, 2003; Ernst & Steinhauser, 2012), sometimes because there are not enough trials of positive feedback. Arbel & Wu (2016) assessed both types of feedback but only found a correlation with learning outcome for the P3a to negative feedback.

Additionally, this delayed recognition result of the present study contrasts that of Butterfield and Mangels (2003), who report a correlation of P3a amplitude with immediate retest performance, but not with delayed retest (also one week). Our finding thus challenges their conclusion that the P300 is not related to memory consolidation. If this holds true in future studies with more participants, this would, to our knowledge, be the first evidence suggesting that the P300 to feedback might not only be involved in working memory and encoding, but also have a long-term memory effect. It is possible, that this was not found by Butterfield and Mangels (2003) because they only looked at negative feedback. We suggest that this long-term effect of a higher P3a amplitude might reflect more attention to positive feedback resulting in the strengthening of already learned information.

No strong evidence was found for a difference between conditions regarding the subsequent memory or subsequent correction effect of the P300. One exception is that the P3a amplitude correlation with delayed recognition accuracy was only significant for the sorted condition. It is possible that the effect is still present in the shuffled condition and was too small to reach significance in the current sample, but it does not appear so based on a mean amplitude plot (see Appendix C, Figure C3C). The findings indicate a stronger effect of P3a amplitude on learning in the sorted condition, possibly due to increased attention to detail.

As mentioned above, there might be more differences in the P300 amplitude that we were not able to uncover in the small sample. A future study might find a larger P3a in the sorted condition, as weakly suggested by visual inspections of the signals and by the finding regarding an increase in the P3a amplitudes to negative feedback. Taken together with the association of P3a amplitude and successful learning, an enhanced P3a or enhanced P3a benefit could potentially be involved in the learning advantage of the sorted condition. Further, as we did find an effect of condition on the P3b amplitude, this could also point towards a beneficial feedback processing difference, since other studies do report a higher P3b amplitude to be correlated with better learning outcomes (e.g., Bultena et al., 2017; Ernst & Steinhauser, 2012).

### *The FRN and feedback utilization*

Although the results show a correlation between FRN amplitude and learning, it is difficult to identify why a *smaller* FRN seems to be related to subsequent correction and better posttest performance. Two previous studies also report a similar effect: Ernst and Steinhauser (2012) found a lower FRN amplitude to subsequently corrected negative feedback

trials, and Arbel and Wu (2016) reported a smaller FRN to negative feedback correlated with learning outcomes. Ernst and Steinhauser (2012) concluded that learning on their task did not depend on the FRN, whereas Arbel and Wu (2016) view their finding as support of the utilization theory.

Since the FRN overlaps with the P300 (San Martín, 2012), it is possible that a smaller negativity associated with subsequent correction of an error is simply due to a larger positivity of the P3a, which is also associated with subsequent correction and subsequent memory. This is especially concerning since in the present study both associations of a lower FRN with learning are accompanied by a larger P3a. When inspecting the grand average of the signal, in both instances the P3a might have an influence on the FRN differences (especially for the delayed recognition effect), but there also seems to be a difference in the FRN trough independently of the P3a. Although this possibility of the FRN difference being due to a superimposed P3a effect cannot be rejected on the basis of this data, our results do suggest that the FRN amplitude plays a role in the effectiveness of feedback processing in our task.

Previous research suggests that the feedback utilization theory applies to reinforcement effects of valence feedback (Ernst & Steinhauser, 2012) and not to the encoding of corrective feedback (Butterfield & Mangels, 2003). In the studies by Arbel and colleagues (Arbel et al., 2014; Arbel et al., 2017; Arbel & Wu, 2016), the subjects received only valence feedback (indicating the correctness of the response). This was possible because they used a two-choice task, and the valence feedback alone was thereby enough for the subjects to learn the correct associations, as opposed to tasks with more choices. This could explain why no learning effects of the FRN are found in most other declarative learning studies (e.g., Bultena et al., 2017; Butterfield & Mangels, 2003; Muller-Gass et al., 2019). In tasks that also include corrective feedback, the informational value of the valence feedback might be relatively less relevant to the learner, resulting in undetectable FRN effects.

As expected, the FRN correlation with subsequent correction and with accuracy on the delayed recognition test were only found in the sorted condition (H3c and H3d). It is again possible that it did not reach significance in the shuffled condition, due to the small sample size. Based on the current data, however, this effect could suggest that the FRN difference – potentially reflecting more effective processing of the feedback –, is only present or stronger in the sorted condition. In the sorted condition, learning might depend more on reinforcement learning processes involving the utilization of valence feedback, because conscious processing and encoding of the informative feedback is more difficult and confusing. The increased utilization of valence feedback might then contribute to an enhanced recall on the posttests, even if no immediate effect – in terms of performance during the task – is visible in the behavioral data.

A possible explanation for the directionality of the FRN correlation with learning would be the account by Arbel and Wu (2016) proposing that faster learners use less processing resources to extract information from negative feedback, which leads to a smaller FRN amplitude as learning progresses. In light of this theory, our finding of a smaller FRN amplitude associated with subsequent correction and subsequent memory could be interpreted as more efficient processing of feedback. Possibly, we observed these correlates

of efficient processing only in the sorted condition because feedback in the sorted condition provides more specific information to the learners, since it reflects smaller orthographic mistakes in the case of negative feedback. The findings of the present study could thus be viewed as an additional piece of evidence suggesting that the FRN reflects not only the processing of valence feedback but also its utilization for learning.

### Limitations

In general, the most prominent limitation of our study was doubtlessly the small sample size of two subjects. This not only affected the statistical significance and value of the findings, but on top of that decreased the quality of the data. The EEG signal of one of the subjects was rather noisy, and this subject would have likely been excluded from the ERP analysis under different circumstances. Moreover, our signal was already noisy overall. With the mobile EEG system that was used, we experienced more high-frequency noise than would be expected in a laboratory. Further, it should be noted that the p-value method that was used (Satterthwaite approximation) is thought to be rather anti-conservative, especially with small sample sizes. Additionally, many models were computed (partially to avoid collinearity of fixed effects) without a correction of multiple comparison. Even so, effects are often found only at one electrode, which could also be due to the small number of participants. Further, the estimates calculated by the models are not very robust under these circumstances. In a few cases of models including interaction effects, the estimates are very different from what the visualized data suggests, likely due to the multiple factors influencing them, but also to the small sample size. The small number of participants also meant that mixed models with random intercepts for participants only had two levels for this factor, which can cause instability in the model. Further, the small sample size made it difficult to evaluate whether the model assumptions were met. Taken together, the statistical results of this study should thus be taken with a grain of salt and seen more as an inspiration for potential future studies.

Another important point is the measurement ambiguity of the FRN component. Because of the temporal overlap between the FRN and the P300, as well as latency differences between participants, the FRN is often assessed as a peak-to-trough measurement (Bultena et al., 2017). In the current study however, a mean amplitude approach was used because it seemed a more robust measure when extracting single-trial ERPs. Further, the latencies of both participants were so different in our study (see Appendix E) that even with a peak-to-trough measure it would have been impossible to pick a time window that would ensure the right trough and peak to be found in both participants. We therefore decided that choosing a slightly larger time-window for mean-amplitude extraction would be the best option for capturing the FRN. Future studies should test whether the FRN effect on learning success can be replicated with a peak-to-trough measure. In addition to the latency difference between participants there was also a substantial difference in FRN latency between Fz and Cz electrodes. Further, the identification of the FRN component was not trivial in the present sample, as there were two negative peaks in the typical time window of 200–300 ms post feedback. Taken together, this unfortunately makes a comparison of the present study to others less straight forward.

Lastly, in light of heightened attention being related to a larger P3a as well as possibly involved in contrasting (Llompert & Reinisch, 2020), it would have been interesting to assess attention via a time-frequency analysis in addition to the ERP analysis. Unfortunately, this was not possible due to a technical difficulty in applying a high-pass filter to the data.

### Future directions and relevance

This study can be seen as a pilot providing some first tentative evidence for a potential benefit of similarity training on feedback processing. In the future it will first of all be interesting to see if the results can be replicated with a larger sample, and whether possibly more robust evidence for a difference between the sorted and the shuffled condition can be found in terms of ERP components. Further, it would be interesting to see if a larger sample would reproduce our findings of ERP correlations with learning being mostly only significant for the sorted condition. In terms of the P300 it should be examined further in which direction the relationship between its amplitude and the similarity manipulation goes and whether this might be influenced by individual performance or by some control variables (e.g., general language learning abilities). Especially in terms of the FRN it is possible that its correlation with learning is not always found because the FRN effects might be obscured by less experienced or less successful learners. The study by Arbel et al. (2014), for example found a correlation with learning outcomes by investigating individual differences, while a study by Themanson et al. (2019) reported that FRN (learning) effects were only found in expert learners. Further investigations are needed to assess why a smaller FRN amplitude was associated with learning in this and previous studies. Since we also proposed that the differences between conditions might be due to expectancy or certainty, future studies might also assess confidence regarding responses. To investigate whether attention to detail is indeed enhanced by contrasting and related to learning success, an assessment of time-frequency parameters or eye-tracking would be of interest.

Studies such as these cannot only uncover more about general feedback processing mechanisms, but they could also in the future be used in the evaluation and improvement of effectiveness of learning paradigms and tools. The ERP components investigated and discussed in this study are well suited for this, because differences can be observed before they become behaviorally visible. For example, the benefit of the similar answer options was only visible in the posttests, but not during the learning task itself, whereas the ERP differences were observed during the task. To our knowledge there has only been one study that investigated potential practical applications in education. This study (Anderson et al., 2018) yielded promising results, showing that feedback-related ERPs can be useful for tracking learning and feedback utilization in pre-classroom learning tools. Since these ERP components can even be used to detect individual differences in the feedback processing effectiveness, such studies could be used to evaluate and create individualized learning support.

### Conclusion

In the present study contrasting similar words during a multiple-choice vocabulary learning task led to slower

learning but improved performance at immediate and delayed posttests. A comparison of ERPs between the sorted and shuffled condition revealed a higher P3b amplitude in the sorted condition, suggesting improved memory processes to be associated with contrasting, although the P3b was not associated with performance in this study. The results on the P3a support its reflection of an evaluation of the response and memory updating based on feedback. Further, this study provides first tentative evidence that a higher P3a to positive feedback might be related to better long-term memory encoding. The ERP results also provide support for the theory of the FRN being related to the utilization of valence feedback for learning. This effect was only found in the condition of similar answer options, suggesting more effective processing of valence feedback when the task involves contrasting similar words. Overall, although we found only limited evidence suggesting that contrasting similar items influences feedback processing, the results suggest that a study of larger sample size and a better controlled experimental environment might still uncover an association between the benefit of contrasting and the effectiveness of feedback processing.

## ACKNOWLEDGMENTS

We would like to thank the technical support group of the faculty of social science at Radboud University, especially Pascal de Water for checking the experiment scripts, and Gerard van Oijen for building the custom made button box (which will be useful in a future replication of the study on campus). Thanks go to Ilaria Lisi for proof-reading the Italian stimuli. Further, we would like to thank Prof. Dr. Peter Desain for lending us the mobile EEG system, as well as Ceci Verbaarschot and Philip van den Broek for providing a sample fieldtrip script and helping with preprocessing questions. We also thank Dr. Duru Özkan for substantial help with the preprocessing scripts. We are grateful to our colleagues of the Social Educational Neuroscience and Artificial Intelligence lab group at Donders Institute for helpful feedback and input about the study design. Thanks also go to Larissa Samaan, Lucas Wolfsturm, and Roman Fränkel for proof-reading the manuscript.

Finally, I (LW) would like to express my gratitude towards Randi Goertz, Peta Baxter, and Dr. Frank Léoné for supervising this internship and master thesis, and for all the support and help with the study design, structuring, the manuscript, and personal goals. Additionally, thanks go to Randi Goertz for data collection and to Peta Baxter for providing example files of scripts and stimuli.

## References

- Allaire J, Xie Y, McPherson J, Luraschi J, Ushey K, Atkins A, Wickham H, Cheng J, Chang W, Iannone R (2020). rmarkdown: Dynamic Documents for R. [R package version 2.3]. <https://github.com/rstudio/rmarkdown>.
- Anderson, S. J., Hecker, K. G., Krigolson, O. E., & Jamniczky, H. A. (2018). A reinforcement-based learning paradigm increases anatomical learning and retention—A neuroeducation study. *Frontiers in Human Neuroscience*, 12(2), 1–10. <https://doi.org/10.3389/fnhum.2018.00038>
- Arbel, Y., Hong, L., Baker, T. E., & Holroyd, C. B. (2017). It's all about timing: An electrophysiological examination of feedback-based learning with immediate and delayed feedback. *Neuropsychologia*, 99(9), 179–186. <https://doi.org/10.1016/j.neuropsychologia.2017.03.003>
- Arbel, Y., Murphy, A., & Donchin, E. (2014). On the utility of positive and negative feedback in a paired-associate learning task. *Journal of cognitive neuroscience*, 26(7), <https://doi.org/1445-1453.10.1162/jocn.a.00617>
- Arbel, Y., & Wu, H. (2016). A Neurophysiological examination of quality of learning in a feedback-based learning task. *Neuropsychologia*, 93(9), 13–20. <https://doi.org/10.1016/j.neuropsychologia.2016.10.001>
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, 68(3), 255–278. <http://dx.doi.org/10.1016/j.jml.2012.11.001>
- Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1). <https://doi.org/10.18637/jss.v067.i01>
- Baxter P.A. et al. (2020). [Title to come and full author list to come]. Manuscript in preparation. Donders Institute for Brain, Cognition, and Behavior. Radboud University.
- Bellebaum, C., & Daum, I. (2008). Learning-related changes in reward expectancy are reflected in the feedback-related negativity. *European Journal of Neuroscience*, 27(7), 1823–1835. <https://doi.org/10.1111/j.1460-9568.2008.06138.x>
- Bellmund, J. L. S., Gärdenfors, P., Moser, E. I., & Doeller, C. F. (2018). Navigating cognition: Spatial codes for human thinking. *Science* (New York, N.Y.), 362(6415) 1–11. <https://doi.org/10.1126/science.aat6766>
- BrainVision Analyzer (Vers. 2.2.0) [Computer Software]. (2020). Gilching, Germany: Brain Products GmbH.
- Bultena, S., Danielmeier, C., Bekkering, H., & Lemhöfer, K. (2017). Electrophysiological correlates of error monitoring and feedback processing in second language learning. *Frontiers in Human Neuroscience*, 11, 29. <https://doi.org/10.3389/fnhum.2017.00029>
- Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods and Research*, 33(2), 261–304. <https://doi.org/10.1177/0049124104268644>
- Butterfield, B., & Mangels, J. A. (2003). Neural correlates of error detection and correction in a semantic retrieval task. *Cognitive Brain Research*, 17(3), 793–817. [https://doi.org/10.1016/S0926-6410\(03\)00203-9](https://doi.org/10.1016/S0926-6410(03)00203-9)
- Chase, H. W., Swainson, R., Durham, L., Benham, L., & Cools, R. (2011). Feedback-related negativity codes prediction error but not behavioral adjustment during probabilistic reversal learning. *Journal of Cognitive Neuroscience*, 23(4), 936–946. <https://doi.org/10.1162/jocn.2010.21456>
- Cohen, M. X., & Ranganath, C. (2007). Reinforcement learning signals predict future decisions. *Journal of Neuroscience*, 27(2), 371–378. <https://doi.org/10.1523/JNEUROSCI.4421-06.2007>
- Ernst, B., & Steinhauser, M. (2012). Feedback-related brain activity predicts learning from feedback in multiple-choice testing. *Cognitive, Affective and Behavioral Neuroscience*,

- 12(2), 323–336. <https://doi.org/10.3758/s13415-012-0087-9>
- Frömer, R., Maier, M., & Rahman, R. A. (2018). Group-level EEG-processing pipeline for flexible single trial-based analyses including linear mixed models. *Frontiers in Neuroscience*, 12(2), 1–15. <https://doi.org/10.3389/fnins.2018.00048>
- Hajcak, G., Moser, J. S., Holroyd, C. B., & Simons, R. F. (2007). It's worse than you thought: The feedback negativity and violations of reward prediction in gambling tasks. *Psychophysiology*, 44(6), 905–912. <https://doi.org/10.1111/j.1469-8986.2007.00567.x>
- Hart, L. A. (2006). A training study using an artificial orthography: Effects of reading experience, lexical quality, and text comprehension in L1 and L2. *Dissertation Abstracts International, B: Sciences and Engineering*, 66(9), 5115.
- Hart, L., & Perfetti, C. A. (2008). Learning words in Zekkish: Implications for understanding lexical representation. *Single word reading: Behavioral and biological perspectives*, 107-128.
- Hattie, J. (1999). Influences on student learning. 1–29. University of Auckland. Retrieved from <http://growthmindseteaz.org/johnhattie.html>
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. <https://doi.org/10.3102/003465430298487>
- Heldmann, M., Rüsseler, J., & Münte, T. F. (2008). Internal and external information in error processing. *BMC Neuroscience*, 9(33), . <https://doi.org/10.1186/1471-2202-9-33>
- Hillyard, S. A. (1985). Electrophysiology of human selective attention. *Trends in Neurosciences*, 8, 400-405. [https://doi.org/10.1016/0166-2236\(85\)90142-0](https://doi.org/10.1016/0166-2236(85)90142-0)
- Holroyd, C. B., Hajcak, G., & Larsen, J. T. (2006). The good, the bad and the neutral: Electrophysiological responses to feedback stimuli. *Brain Research*, 1105(1), 93–101. <https://doi.org/10.1016/j.brainres.2005.12.015>
- Holroyd, C. B., & Krigolson, O. E. (2007). Reward prediction error signals associated with a modified time estimation task. *Psychophysiology*, 44(6), 913–917. <https://doi.org/10.1111/j.1469-8986.2007.00561.x>
- Huelser, B. J., & Metcalfe, J. (2012). Making related errors facilitates learning, but learners do not know it. *Memory and Cognition*, 40(4), 514–527. <https://doi.org/10.3758/s13421-011-0167-z>
- Janssen, C., Segers, E., McQueen, J. M., & Verhoeven, L. (2015). Lexical specificity training effects in second language learners. *Language Learning*, 65(2), 358–389. <https://doi.org/10.1111/lang.12102>
- Johnson, R. (1986). A triarchic model of P300 amplitude. *Psychophysiology*, 23(4), 367-384.
- Jongsma, M. L. A., Gerrits, N. J. H. M., van Rijn, C. M., Quiroga, R. Q., & Maes, J. H. R. (2012). Event related potentials to digit learning: Tracking neurophysiologic changes accompanying recall performance. *International Journal of Psychophysiology*, 85(1), 41–48. <https://doi.org/10.1016/j.ijpsycho.2011.10.004>
- Karis, D., Fabiani, M., & Donchin, E. (1984). “P300” and memory: Individual differences in the von Restorff effect. *Cognitive Psychology*, 16(2), 177–216. [https://doi.org/10.1016/0010-0285\(84\)90007-0](https://doi.org/10.1016/0010-0285(84)90007-0)
- Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, 42(3), 627–633. <https://doi.org/10.3758/BRM.42.3.627>
- Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, 42(3), 643–650. <https://doi.org/10.3758/BRM.42.3.643>
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of statistical software*, 82(13), 1-26. <https://doi.org/10.18637/jss.v082.i13>
- Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, 10 (8), 707-710.
- Little, J. L., & Bjork, E. L. (2015). Optimizing multiple-choice tests as tools for learning. *Memory & Cognition*, 43, 14–26. <https://doi.org/10.3758/s13421-014-0452-8>
- Llach, P.A. (2015). Lexical Errors in Writing at the End of Primary and Secondary Education : Description and Pedagogical Implications. *Porta Linguarum*, 23, 109–124
- Llompert, M., & Reinisch, E. (2020). The phonological form of lexical items modulates the encoding of challenging second-language sound contrasts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46(8), 1590–1610. <https://doi.org/10.1037/xlm0000832>
- Ludowicy, P., Czernochowski, D., Weis, T., Haese, A., & Lachmann, T. (2019). Neural correlates of feedback processing during a sensory uncertain speech – nonspeech discrimination task. *Biological Psychology*, 144(7), 103–114. <https://doi.org/10.1016/j.biopsycho.2019.03.017>
- Luu, P., Tucker, D. M., Derryberry, D., Reed, M., & Poulsen, C. (2003). Electrophysiological responses to errors and feedback in the process of action regulation. *Psychological Science*, 14(1), 47–53. <https://doi.org/10.1111/1467-9280.01417>
- MATLAB. (2020). Version 9.8 (R2020a). [Computer software]. Natick, Massachusetts: The MathWorks Inc.
- Metcalfe, J. (2017). Learning from Errors. *Annual Review of Psychology*, 68(1), 465–489. <https://doi.org/10.1146/annurev-psych-010416-044022>
- Meteyard, L., & Davies, R. A. I. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language*, 112(2), 1-22. <https://doi.org/10.1016/j.jml.2020.104092>
- Muller-Gass, A., Duncan, M., Tavakoli, P., & Campbell, K. (2019). Individual differences in feedback processing affect perceptual learning. *Personality and Individual Differences*, 143(10), 145–154. <https://doi.org/10.1016/j.paid.2019.01.017>
- Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. M. (2011). FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational intelligence and neuroscience*, 2011, 156869. <https://doi.org/10.1155/2011/156869>
- Peirce, J. W. (2007). PsychoPy-Psychophysics software in Python. *Journal of Neuroscience Methods*, 162(1–2), 8–13. <https://doi.org/10.1016/j.jneumeth.2006.11.017>
- Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, 11(4), 357–383. <https://doi.org/10.1080/10888430701530730>
- Philiastides, M. G., Biele, G., Vavatzanidis, N., Kazzner, P., & Heekeren, H. R. (2010). Temporal dynamics of prediction error processing during reward-based decision making.

- NeuroImage*, 53(1), 221–232.  
<https://doi.org/10.1016/j.neuroimage.2010.05.052>
- Polich, J. (2007). Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology*, 118(10), 2128–2148.  
<https://doi.org/10.1016/j.clinph.2007.04.019>
- R Core Team (2020). R: A language and environment for statistical computing [Computer software]. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
- Reder, L. M., Liu, X. L., Keinath, A., & Popov, V. (2016). Building knowledge requires bricks, not sand: The critical role of familiar constituents in learning. *Psychonomic Bulletin and Review*, 23(1), 271–277.  
<https://doi.org/10.3758/s13423-015-0889-1>
- San Martín, R. (2012). Event-related potential studies of outcome processing and feedback-guided learning. *Frontiers in Human Neuroscience*, 6(10), 1–40.  
<https://doi.org/10.3389/fnhum.2012.00304>
- Sanquist, T. F., Rohrbaugh, J. W., Syndulko, K., and Lindsay, D. B. (1980). Electrocardinal signs of levels of processing: perceptual analysis and recognition memory. *Psychophysiology*, 17, 568–576.  
<https://doi.org/10.1111/j.1469-8986.1980.tb02299.x>
- Steinemann, N. A., Moisello, C., Ghilardi, M. F., & Kelly, S. P. (2016). Tracking neural correlates of successful learning over repeated sequence observations. *NeuroImage*, 137, 152–164.  
<https://doi.org/10.1016/j.neuroimage.2016.05.001>
- Themanson, J. R., Bing, N. J., Sheese, B. E., & Pontifex, M. B. (2019). The influence of pitch-by-pitch feedback on neural activity and pitch perception in baseball. *Journal of Sport and Exercise Psychology*, 41(2), 65–72.  
<https://doi.org/10.1123/jsep.2018-0165>
- Van Casteren, M., & Davis, M. H. (2006). Mix, a program for pseudorandomization. *Behavior Research Methods*, 38(4), 584–589. <https://doi.org/10.3758/BF03193889>
- van de Ven, M., Segers, E., & Verhoeven, L. (2019). Enhanced Second Language Vocabulary Learning Through Phonological Specificity Training in Adolescents. *Language Learning*, 69(1), 222–250.  
<https://doi.org/10.1111/lang.12330>
- van der Helden, J., Boksem, M. A. S., & Blom, J. H. G. (2010). The importance of failure: Feedback-related negativity predicts motor learning efficiency. *Cerebral Cortex*, 20(7), 1596–1603. <https://doi.org/10.1093/cercor/bhp224>
- Walsh, M. M., & Anderson, J. R. (2012). Learning from experience: Event-related potential correlates of reward processing, neural adaptation, and behavioral choice. *Neuroscience and Biobehavioral Reviews*, 36(8), 1870–1884. <https://doi.org/10.1016/j.neubiorev.2012.05.008>
- Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications (Cognitive Information Science). 1–42. University of California.  
<https://doi.org/http://arxiv.org/pdf/1308.5499.pdf>

## APPENDIX A

### Description of the Original Design / Plan for a Future Continuation of the Project

#### Participants

All subjects will be native speakers of Dutch. The participants will be recruited online via the Radboud Research Participation System (SONA) and receive course credits or 10€ per hour.

#### Stimuli

The stimulus material consists of 50 Dutch words and 50 pseudo-words, each as a translation for one of the Dutch words. The Dutch words are concrete nouns, four to six letters long, and their frequency (per million in SUBTLEX-NL; Keuleers, Brysbaert, & New, 2010) was between 0.5 and 65.13 ( $M = 19.11$ ). None of the Dutch words are semantically or orthographically similar to each other (minimum Levenshtein distance: 2). The Pseudo-words are the same as in the Italian-word study, except for one sorted group (“abid”, “abet”, “abec”, “abic”, “abit”; replaced by “gero”, “veco”, “vero”, “vemo”, “gemo”) and one additional word (“mulir”; replaced by “mugir”). The Levenshtein distance between words within a sorted-group is between 1 and 2, whereas the Levenshtein distance between words of different sorted-groups is minimally 3. The similarity manipulation is implemented in the same way as described in the paper.

#### Design and procedure

Testing will take place in the lab. There will be two sessions, as described above. During the tasks, the participants are seated in front of a computer monitor with 1920x1080 resolution. For the learning task and the translation recognition task, an in-house designed (<https://www.ru.nl/socialsciences/technicalsupportgroup/>) button box with three horizontally aligned buttons will be used. During the translation production task, the subjects respond by typing on a computer keyboard. The learning task and the posttests are the same as described in the paper.

#### EEG measurement

EEG signals will be recorded during the learning task at a sampling rate of 500Hz. Thirty-two active electrodes will be mounted on an elastic cap according to a 10–20 configuration. Additionally, electrooculogram (EOG) will be recorded with four electrodes from below and above the right eye and on the temples. Electrodes will be online averaged to left mastoid. A low cutoff filter of 10 s (frequency 0.016 Hz) and high cutoff filter of 125 Hz will be used.

#### Data analysis

The EEG signal will be preprocessed in BrainVision Analyzer. The statistical analysis will be carried out in the same way as described above.

## APPENDIX B

### Translation recognition task: Selection procedure of translation pairs

For a participant seeing set I in the sorted condition, one entire sorted-group of set I was randomly selected to be presented as correct pairs. Further, from each of the remaining four sorted-groups, one additional word was pseudo-randomly selected to be presented as a correct pair. The same was done for set II, but with the shuffled-groups.

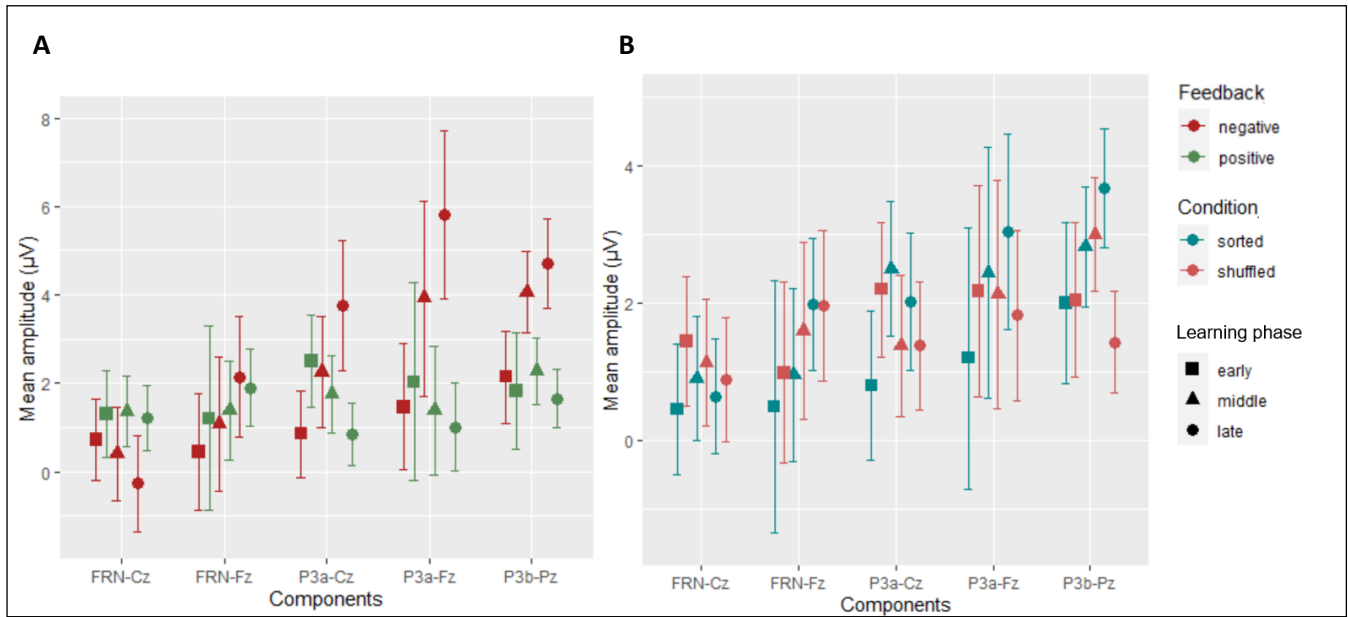
For the remaining 16 words in set I, in each sorted-group two Italian words were selected, and their pseudo-translations were exchanged. This resulted in eight Italian words that were presented with an incorrect pseudo-translation that was similar to the correct translation *and* had been presented as an alternative answer option of that Italian word during the task. The remaining eight Italian words of set I were pseudo-randomly paired with a pseudo-word from another sorted-group of set I – a word their correct translation was neither similar to nor presented with during the learning task.

For set II, two Italian words of each of the four remaining shuffled-groups were selected and their pseudo-translations exchanged, resulting in eight words paired with an option they were not similar to, but one that had been presented as an alternative option during the learning task. The remaining eight Italian words of set II were paired with pseudo-words from other shuffled-groups, but from the same similar-groups – such that they would be similar to the correct translation, but had not been presented as alternative options of the given Italian word before.

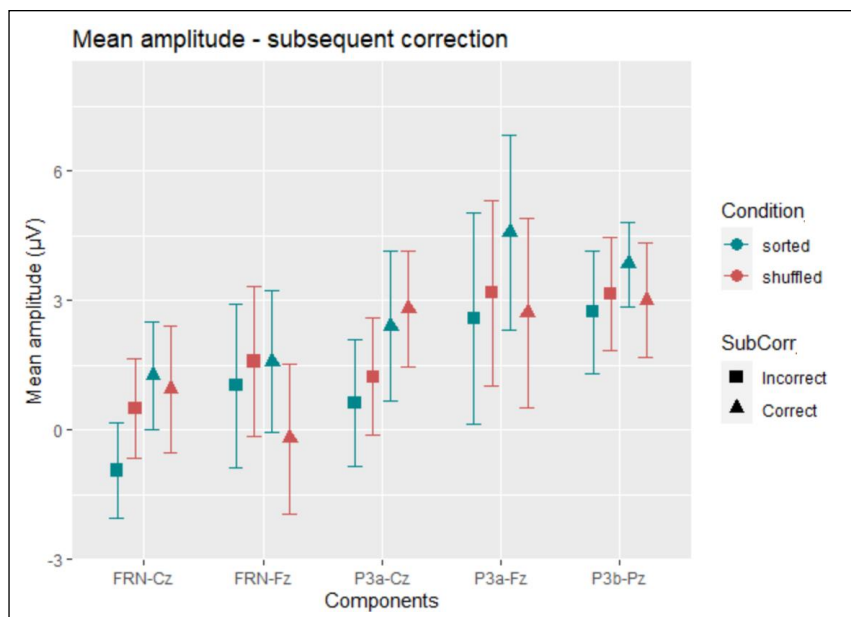
Overall, this resulted in nine correctly paired words in the shuffled condition, and nine correctly paired in the sorted condition. Further, there were eight paired with a translation similar to the correct target that were an alternative option during the learning task. Eight words were paired with a translation that was dissimilar to the correct target and that had never been presented as an alternative option for that word. Another eight were paired with a pseudo-word that was dissimilar to the correct translation but that had been presented as an alternative option. Lastly, eight words were paired with a pseudo-word dissimilar to the correct translation and that had not been presented as an option for this word.

## APPENDIX C

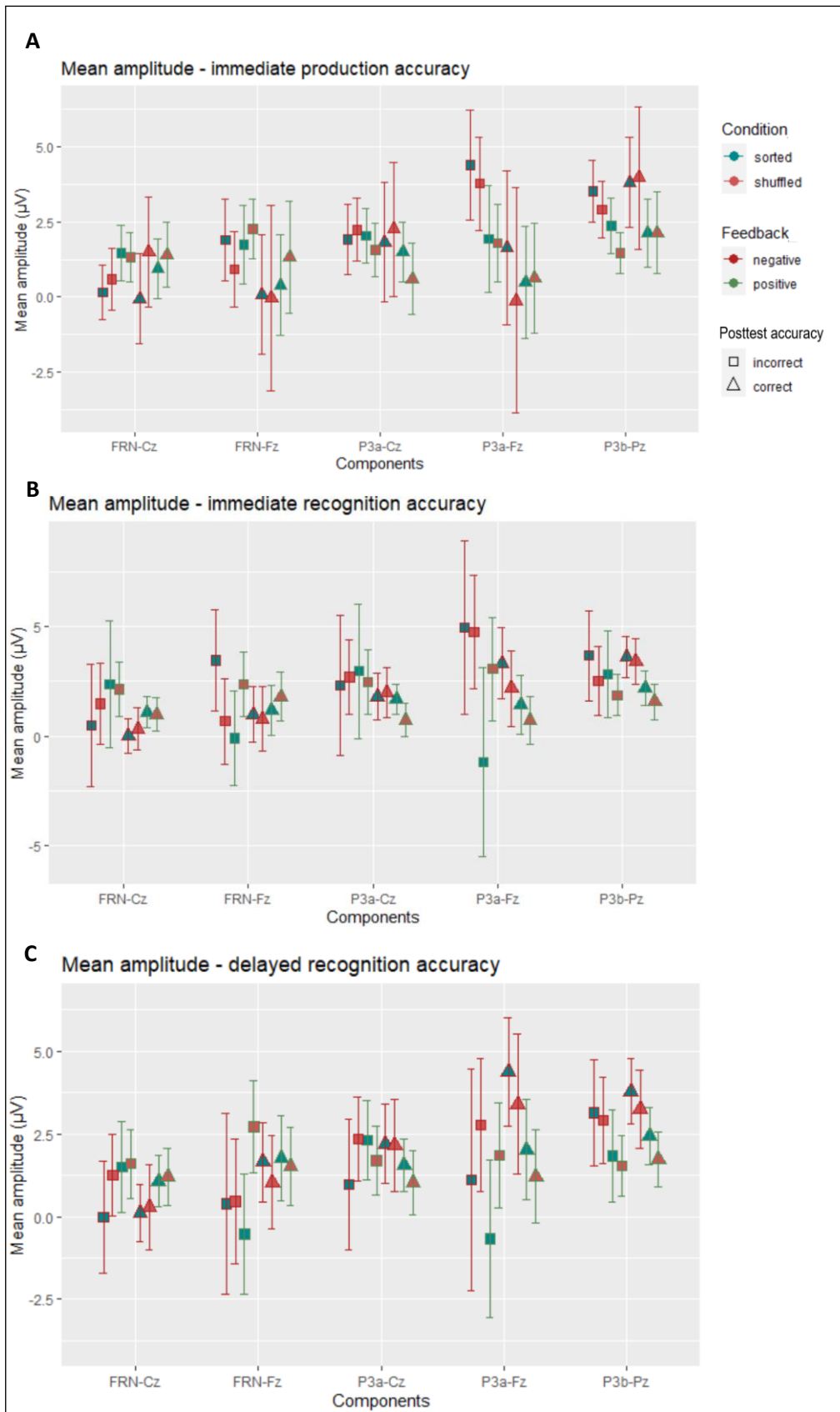
Mean ERP amplitudes of each component grouped by fixed effects



**Figure C1.** A: Mean amplitudes grouped by learning phase and feedback type. B: Mean amplitudes grouped by condition and learning phase.



**Figure C2.** Mean amplitudes grouped by condition and subsequent correction.



**Figure C3.** A: Mean amplitudes grouped by immediate production accuracy, condition, and feedback type. B: Mean amplitude grouped by immediate recognition accuracy, condition, and feedback type. C: 1. Mean amplitude grouped by delayed recognition accuracy, condition, and feedback type.

**APPENDIX D**

Results of immediate production and recognition models

**Table D1.** Immediate production model

<b>P300</b>		<b>P3a Fz</b>				<b>P3a Cz</b>				<b>P3b Pz</b>					
<b>Fixed effects</b>	<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>		<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>	<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>		
Intercept	3.33	1.75	1.91	.175		1.93	0.89	2.17	.128	3.72	0.63	5.94	.006	**	
FB_pos-neg	-1.50	0.65	-2.32	.021	*	-0.35	0.42	-0.82	.407	-1.50	0.39	-3.85	<.001	***	
Cond_Sh-So	-0.30	0.64	-0.46	.646		-0.19	0.42	-0.46	.647	-0.53	0.39	-1.38	.169		
IP_C-I	-1.02	0.70	-1.46	.144		-0.03	0.45	-0.06	.950	0.06	0.42	0.14	.890		
<b>Random Effects</b>	<b>Variance</b>				<b>SD</b>	<b>Variance</b>				<b>SD</b>	<b>Variance</b>				<b>SD</b>
Participant (Intercept)	5.37				2.32	1.28				1.13	0.52				0.72
Residual	51.42				7.17	21.58				4.65	18.62				4.31

<b>FRN</b>		<b>FRN Fz</b>				<b>FRN Cz</b>				
<b>Fixed effects</b>	<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>		<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>	
Intercept	1.03	1.01	1.02	.382		-0.14	0.97	-0.14	.899	
FB_pos-neg	0.12	0.54	0.23	.817		0.99	0.37	2.64	.008	**
Cond_Sh-So	0.75	0.54	1.37	.169		0.39	0.37	1.06	.288	
IP_C-I	-0.81	0.58	-1.39	.165		0.46	0.40	1.15	.249	
<b>Random Effects</b>	<b>Variance</b>				<b>SD</b>	<b>Variance</b>				<b>SD</b>
Participant (Intercept)	1.52				1.23	1.65				1.29
Residual	35.80				5.98	16.92				4.11

FB, Feedback: pos = positive, neg: negative; Cond, Condition: Sh = shuffled, So = sorted; IP, immediate production: C = correct response, I = incorrect response  
 \*\*\*p < .001, \*\*p < .01, \*p < .05

**Table D2.** Immediate recognition model

<b>P300</b>		<b>P3a Fz</b>				<b>P3a Cz</b>				<b>P3b Pz</b>					
<b>Fixed effects</b>	<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>		<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>	<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>		
Intercept	4.82	2.57	1.88	.154		2.24	1.17	1.92	.094	3.27	0.88	3.73	.002	**	
Cond_Sh-So	-1.23	1.03	-1.19	.236		0.12	0.74	0.16	.870	-0.37	0.65	-0.58	.564		
RI_C-I	-1.92	1.22	-1.58	.116		-0.55	0.87	-0.63	.532	0.47	0.76	0.61	.545		
<b>Random Effects</b>	<b>Variance</b>				<b>SD</b>	<b>Variance</b>				<b>SD</b>	<b>Variance</b>				<b>SD</b>
Participant (Intercept)	10.09				3.18	1.14				1.07	0.32				0.56
Residual	51.98				7.21	26.61				5.16	20.44				4.52

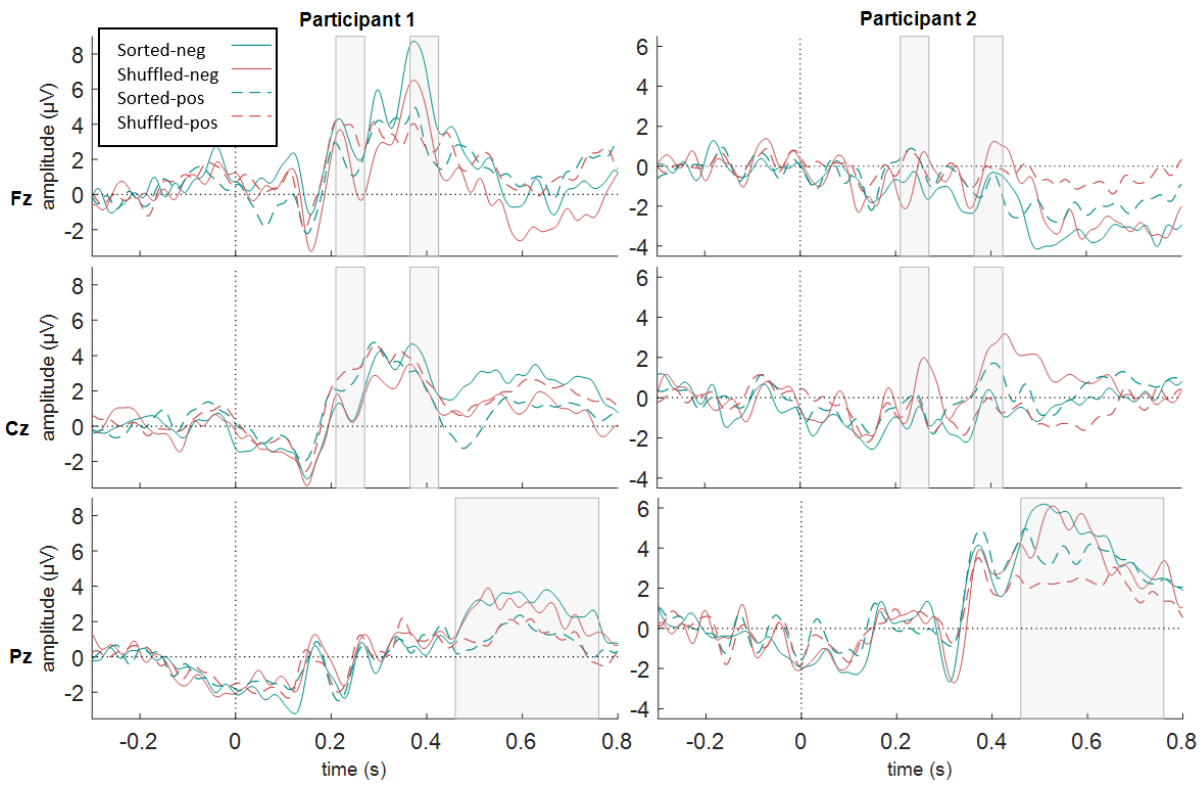
  

<b>FRN</b>		<b>FRN Fz</b>				<b>FRN Cz</b>				
<b>Fixed effects</b>	<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>		<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>	
Intercept	1.83	1.46	1.25	.260		0.76	0.83	0.91	.375	
Cond_Sh-So	-0.90	0.84	-1.07	.286		0.42	0.62	0.69	.493	
RI_C-I	-0.78	1.00	-0.78	.434		-0.83	0.73	-1.15	.253	
<b>Random Effects</b>	<b>Variance</b>				<b>SD</b>	<b>Variance</b>				<b>SD</b>
Participant (Intercept)	2.20				1.48	0.28				0.53
Residual	34.70				5.89	18.51				4.30

Cond, Condition: Sh = shuffled, So = sorted; RI, immediate recognition: C = correct response, I = incorrect response  
 \*\*\*p < .001, \*\*p < .01, \*p < .05

## APPENDIX E

### Individual mean amplitude plots



**Figure E.** Mean ERP amplitudes depicted separately for the two subjects, comparing positive and negative feedback, as well as sorted and shuffled conditions. Note the differential scaling of the y-axis per participant.