Automatically Extracting ERP Component Latencies Using a Template Matching Algorithm

Sven Lesche1

1 Ruprecht-Karls-University Heidelberg

Author note

Author Notes go here.

The authors made the following contributions. Sven Lesche: Conceptualization, Writing - Original Draft Preparation, Writing - Review & Editing.

Correspondence concerning this article should be addressed to Sven Lesche, Im Neuenheimer Feld 695, 69120 Heidelberg. E-mail: [sven.lesche@psychologie.uni-heidelberg.de](mailto:sven.lesche@psychologie.uni-heidelberg.de)

Abstract

One or two sentences providing a **basic introduction** to the field, comprehensible to a scientist in any discipline.

Two to three sentences of **more detailed background**, comprehensible to scientists in related disciplines.

One sentence clearly stating the **general problem** being addressed by this particular study.

One sentence summarizing the main result (with the words “**here we show**” or their equivalent).

Two or three sentences explaining what the **main result** reveals in direct comparison to what was thought to be the case previously, or how the main result adds to previous knowledge.

One or two sentences to put the results into a more **general context**.

Two or three sentences to provide a **broader perspective**, readily comprehensible to a scientist in any discipline.

*Keywords:* keyword, more keyword, crazy keyword

*Word count:* X

Automatically Extracting ERP Component Latencies Using a Template Matching Algorithm

# Introduction

In cognitive neuroscience, the accurate extraction of latencies of event-related potentials (ERPs) stands as a crucial step in understanding the timing of neural processes that underlie cognitive functions ([Luck, 2014](#ref-luck2014introduction); [Meyer et al., 1988](#ref-meyer1988modern); [Posner, 2005](#ref-posner2005timing)). Especially in studies focusing on individual differences an objective, efficient, valid and reliable extraction process is paramount. Algorithms allow efficient and objective approaches to latency extraction, but often prove unreliable and invalid due to ERPs with low signal-to-noise ratios ([Clayson et al., 2013](#ref-clayson2013noise); [Kiesel et al., 2008](#ref-kiesel2008measurement); [Sadus et al., 2023](#ref-sadus2023multiverse); [Schubert et al., 2023](#ref-schubert2023robust)). To deal with this issue, researchers often forgo algorithms and manually inspect each ERP, identifying patterns reflecting the component of interest by hand ([Sadus et al., 2023](#ref-sadus2023multiverse)). Manual extraction is a labor-intensive and time-consuming endeavor, improving reliability and validity at the cost of objectivity and efficiency. In an effort to improve replicability, it is recommendable that all processing steps be automated ([Cowley et al., 2017](#ref-cowley2017computational); [Rodrigues et al., 2021](#ref-rodrigues2021epos)). This paper introduces a novel algorithm for the automatic extraction of ERP latencies using template matching. We aim to show that our algorithm improves on existing approaches and enables more efficient, objective, reliable and valid extraction of ERP latencies.

## Peak Latency Algorithms

Automatically extracting component latencies has long been a goal in ERP research. *Peak latency* algorithms present the earliest attempt at automating the latency extraction process. The peak latency approach involves finding the point in time within a set measurement window that displays the largest voltage deflection in the appropriate direction. This approach remains the most common way of latency extraction ([Kiesel et al., 2008](#ref-kiesel2008measurement); [Liesefeld, 2018](#ref-liesefeld2018estimating)), partially due to the ease of its implementation and low computational requirements ([Donchin & Heffley, 1978](#ref-donchin1978multivariate)). However, ease of implementation is accompanied by several drawbacks.

Firstly, peak latency approaches are blind to the general structure within the ERP signal. They locate the point in time with the largest voltage deflection within the measurement window. The “true” maximum signal may lay just outside of the measurement window, resulting in the algorithm picking an outside edge of the window, even though the signal increases in amplitude just following that time-point. The simple peak latency algorithm can be extended to only consider maxima inside the measurement window which are also larger than the surrounding data-points ([Luck, 2014](#ref-luck2014introduction)). A spike just on the edge of the measurement window will not meet these criteria. This approach protects against the influence of noise or surrounding components on the edges of the measurement window, but remains vulnerable to high frequency noise inducing peaks inside the measurement window.

Due to high frequency noise, the maximum voltage deflection may not reflect the true point in time at which the process underlying the component reaches its maximum. Rather it may be the result of high frequency noise inducing a spike in the signal, independent of any cognitive process. This is problematic especially in later ERP components, as the broader measurement window most commonly applied increases the likelihood of high frequency noise inducing the maximum voltage deflection ([Clayson et al., 2013](#ref-clayson2013noise)).

The sensitivity of the peak latency approach to the size of the measurement window is further increased by the *superimposition problem* ([Luck, 2014](#ref-luck2014introduction)). In larger measurement windows, later components can already influence the amplitude on the outer ranges of the measurement window. Researchers have to carefully choose the measurement window in order to include most of the signal related to the component of interest while simultaneously excluding influences of other components.

Lastly, as Luck ([2014](#ref-luck2014introduction)) so aptly states: *There is nothing special about the point at which the voltage reaches a local maximum*. The largest deflection does not inherently relate to any physiological or psychological process and may not even reflect the true maximum of the component of interest. Luck ([2005](#ref-luck2005ten)) visually demonstrates how peak latency may be a result of the overlap of multiple components and not related to any single component.

## Fractional Area Latency Algorithms

*Fractional Area Latency* approaches hope to remedy some of the problems associated with peak latency algorithms. This technique revolves around the area under the ERP signal in a given measurement window. The goal is to find the point in time that divides the area under the signal into a given fraction to the left and right of it. The time-point halving the area under the signal, for example, is referred to as the *50% area latency*. This approach is much less susceptible to the influence of high frequency noise, as short spikes in the signal do not have a strong impact on the area under the signal ([Liesefeld, 2018](#ref-liesefeld2018estimating)). Nonetheless, area latency approaches remain highly dependent on the measurement window ([Luck, 2014](#ref-luck2014introduction)). Choosing a shorter window may result in only a part of the area of the component of interest being captured. A wider measurement window on the other hand might be influenced by surrounding components. Fractional Area measures thus work best for investigating an isolated component ([Luck, 2014](#ref-luck2014introduction)), limiting the applicability of area latency algorithms.

## Jackknifing

Another approach towards dealing with noisy subject-level ERPs is to try and mitigate that noise by averaging multiple subject-level ERPs. This technique is referred to as *jackknifing* ([Miller et al., 1998](#ref-miller1998jackknife)). It uses the grand average of subject-level ERPs and generates sub-grand averages by removing one of the subject-level ERPs. This averaging procedure results in ERPs with higher signal-to-noise ratios. Both *peak* and *area* based measures can then be applied to the jackknifed data to extract latencies. As any two jackknifed signals share percent of the subject-level signals that are averaged with each other, each jackknifed sub-average is quite similar to all others. This artificially decreases the error variance, which needs to be corrected for when testing for differences between groups ([Ulrich & Miller, 2001](#ref-ulrich2001using)). Because latencies extracted from jackknifed ERPs are based on averaged waveforms, they can not readily be associated with any single subject, preventing this method from generating individual-level latency estimates needed for individual differences research. This problem was addressed by ([Smulders, 2010](#ref-smulders2010simplifying)) who introduced a transformation able to generate individual-level latency estimates.

## Comparison of Algorithms

Kiesel et al. ([2008](#ref-kiesel2008measurement)) applied these different algorithms, as well as a few additional approaches not mentioned here, to a variety of ERP components. They simulated latency differences of the visual and auditory N1, the N2pc and the P3 and frequency-related P3 and tested single-participant approaches and jackknife-approaches combined with peak latency, fractional area latency, relative criteria and baseline deviation methods on their ability to detect these effects. The most widely used technique of single-participant approaches combined with peak latency extraction proved to not be the most best method to detect latency effects. Its effectiveness decreases even further as the signal-to-noise ratio decreases. Overall, jackknifing ERPs and using the relative criterion technique or the fractional area latency technique was shown to be the best approach across the components and datasets they analyzed.

This finding was further corroborated by Wascher et al. ([2022](#ref-wascher2022mental)), who investigated the ability of peak latency and area latency measures combined with jackknifing to generate reliable latency measures. Area latency measures combined with jackknifing generated the most reliable estimates ERP latencies across a variety of components. However, even the best automated approach did not lead to consistently high reliabilities.

Investigating this further, Sadus et al. ([2023](#ref-sadus2023multiverse)) assessed the influence of different preprocessing strategies and latency extraction techniques on psychometric properties of the latency values generated as well as the ability to detect an age-related effect in P3 latency. They varied the strength of the low-pass filter applied to the data, used both single-participant and jackknife approaches and extracted latencies either automatically or manually, using either a peak latency or area latency approach. Both the size of the effect and the psychometric properties, such as reliability or homogeneity of the latency values, varied between the different analysis strategies. No combination of preprocessing steps and extraction method proved best across all tasks and groups. Only 7 out of 40 possible pipelines generated consistently desirable reliabilities (), homogeneities () and effect sizes (). All of those seven pipelines used manual extraction methods either based on peak or area latencies. While automated extraction methods would improve both efficiency and objectivity, fully automated approaches failed to generate consistently reliable and valid latency measures ([Sadus et al., 2023](#ref-sadus2023multiverse); [Schubert et al., 2023](#ref-schubert2023robust)). Yet, manual extraction methods are highly time-consuming and impede reproducibility. We hope to show that our algorithm can match the performance of manual extraction while providing a more efficient and objective approach for extracting individual component latency values.

## Template Matching

The shortcomings of prior algorithms may be due to none of them aiming to replicate human behavior and rather focusing on generating similar decisions, not similar decision processes. Our proposed algorithm more closely resembles the process expert ERP researches employ. During latency extraction, most researchers have some mental representation of what the component *should* look like and where it *should* generally appear. When visually inspecting ERP signals, their goal is to identify a pattern within the signal that resembles the component of interest in shape, size and location.

Finding a given pattern inside a noisy signal is not a novel task. Algorithms aiming to detect the appearance of a pattern, a *template*, inside audio-, video- or radio signals have been around for over 50 years, and a large amount of research has gone into optimizing these *template matching* algorithms ([Briechle & Hanebeck, 2001](#ref-briechle2001template); [Brunelli, 2009](#ref-brunelli2009template); [Lewis, 1995](#ref-lewis1995fast); [Mahalakshmi et al., 2012](#ref-mahalakshmi2012image)).

No matter the implementation details of a particular template matching algorithm, all aim to answer the question “Does this (smaller) template appear in my (larger) signal?”. To achieve this, a researcher needs to specify two things. First, the template they want to search the signal for and second, the *similarity measure* which quantifies how well the template fits in a given spot of the signal.

### Similarity measures.

Specifying the template is mostly a philosophical question dependent on the specific task and type of signal. Choosing a similarity measure on the other hand is much more methodological. Across a number of papers, several different similarity measures have been proposed. They follow one of two general lines of thought ([Brunelli & Poggiot, 1997](#ref-brunelli1997template); [Goshtasby et al., 1984](#ref-goshtasby1984two)). The first type of similarity measure aims to minimize some value reflecting the distance between template and the signal. The second type aims to maximize some form of correlation between signal and template. We have chosen to implement algorithms based on two of the possible measures, one minimizing the sum of squared differences and one maximizing the correlation between the template and the signal. We wanted to implement both a similarity measure following a traditional distance-minimization approach and a correlation-based approach in order to gauge the efficacy of these approaches when applied to our field of research. We call the algorithm based on minimization of the squared differences *MINSQ* and the algorithm based on maximization of the correlation *CORR*.

### Template generation.

Depending on the field of study, the template to search for is easily specified. If you are looking to extract a particular audio-signal from a recording or some specific object in an image you can easily use that object as a template. The difficulty increases if it is not exactly certain what template you are looking for and what shape that template may take depending on various external factors. Recent research in image processing has attempted to use template matching to process faces ([Brunelli, 2009](#ref-brunelli2009template)). You cannot just use “the ideal set of eyes” to identify a face. Each person comes with their own set of eyes, different from all others in some quantifiable way. A similar issue accompanies attempts of template matching approaches in ERP research. The variance in ERP signals introduced by the task or the sample of participants hinders a successful implementation of template matching algorithms using only one idealized template over all types of studies. Finding a template that reflects the influence of the task or the sample, equivalent to finding the person who’s eyes you are looking for, will significantly improve performance.

## Prior template matching algorithms

There have been some attempts at using an idealized signal structure as a template to identify ocular artifacts in noisy subject-level data ([Li et al., 2006](#ref-li2006automatic)) or to predict subject behavior on a single-trial level ([William et al., 2020](#ref-william2020erp)). These studies were able to successfully implement template matching algorithms. However, they were not concerned with estimating the timing of components, but rather only interested in detection of a specific signal.

Borst and Anderson ([2015](#ref-borst2015discovery)) and Anderson et al. ([2016](#ref-anderson2016discovery)) developed a machine-learning approach that aims to discover cognitive processing stages on a single-trial level. In a first step, their algorithm makes use of multivariate pattern analysis to detect “bumps” in the EEG signal representing the onset of a new cognitive state. They assumed that entry into a new state would be accompanied by a bump in all electrodes similar to a 50 ms half-sine. This 50 ms half-sine then serves as a template with which their algorithm tries to detect those “bumps” in activity. However, their assumptions regarding the template and the location of activity are somewhat crude generalizations made necessary by noisy single trial data. Using template matching to extract component latencies from ERPs requires a more informative template than a half-sine.

## Using the grand average as a template

A simple approach towards designing a more informative template would be to generate an idealized component structure. Prior knowledge about the shape, size and location about the component of interest could then be made use of. One could draw up “the perfect P3” and attempt to use this as a template. However, this neglects the experiment-specific and task-specific variance in the morphology of ERP components, resulting in a template that does not optimally reflect the data. We addressed this problem by using the grand average as an experiment-specific template of the component of interest that will reflect influences of the task and sample on the morphology of the idealized ERP component. We use the grand average and the specific window in which the component of interest occurs as a template to match to subject-level ERPs. The grand average has a higher signal-to-noise ratio than any individual subject-level ERP while still being influenced by experiment-specific changes in component morphology. It is by definition the average of all subject-level ERPs and thus minimizes the sum of squared deviations between each subject-level ERP and itself. Thus, across all subjects, it is the best approximation for each subject-level ERP and thus lends itself to being used as a template. In a practical example, researchers can (and already often do) use the grand average to gather insight into the time window in which the component of interest occurs ([Kiesel et al., 2008](#ref-kiesel2008measurement)).

The goal of our algorithm is to quantify individual differences in the latency of ERP components, not only the presence of a particular ERP component. This is similar to facial recognition not only detecting the presence of eyes, but also determining the person the eyes belong to. We do not only aim to detect the presence of a component, but also aim to measure individual differences in shape, size and location of the component.

To achieve this, we introduce variability into the template reflecting individual differences in the morphology of the component. This transformation generates versions of the grand average as a template with different amplitudes and locations of the component of interest. Crudely, each transformation reflects the idealized component if it had higher vs. lower overall amplitudes and earlier vs. later latencies of the component of interest. We can simulate how the template would look like if the component had any given latency. Determining which transformation of the template best fits a given subject-level ERP will allow us to investigate individual differences.

## Mapping individual differences

We introduce variability in the template through two sources, amplitude and latency. Variability in amplitude is controlled by the parameter , variability in latency by the parameter . Amplitude of the template is varied by multiplying the whole template-signal by parameter . Latency is varied by “stretching” or “squishing” the template along the x-axis (see Figure 1). The parameter controls the strength of this transformation. Possible transformations of the template are then compared to the signal and their similarity is evaluated. Recovering the transformation parameters that lead to the best match between template and signal thus allows us to describe individual differences in the latency of a component. To recover latencies of the component of interest from the optimal transformation parameter, researchers need to specify a time-point of the grand average denoting the latency of the component of interest. The optimal transformation parameters are then applied to the grand average latency to recover the subject-level latency of the component.



Figure 1: Scaling Templates Horizontally

## Why our algorithm may perform better

This algorithm aims to address some of the issues faced by other algorithms. It makes use of the entire component structure to construct the template that is matched to subject-level ERPs. This reflects the decision process of expert ERP researchers and enables an intuitive understanding of the decisions made by our algorithm. Because the similarity measures take the whole component structure into account, they are robust to peaks introduced by high frequency noise. Furthermore, the measurement window set by the researcher only impacts the size and shape of the template. It has no direct connection to the subject-level ERP. The influence of measurement windows on the latencies extracted should thus be lower than in peak latency or area latency algorithms. Lastly, it is important to note that this is not a machine learning algorithm with a neural net representing some “black box” decision making algorithm. Simplicity and traceability of the decision process was an express goal, allowing more insight into the benefits and drawbacks of our algorithm.

## The present study

In order to compare the quality of our proposed algorithm we will reanalyze the same data analyzed by Sadus et al. ([2023](#ref-sadus2023multiverse)). We compare the psychometric properties of our algorithm to those of previously established algorithms, investigate the impact of different preprocessing steps and evaluate the correlation between latencies extracted by our algorithm and those extracted manually by an expert ERP researcher.

In their study, Sadus et al. ([2023](#ref-sadus2023multiverse)) extracted latencies of the P3b component, henceforth simply referred to as P3. The P3 is a centro-parietal positive-going component, peaking around 300 ms after stimulus onset. It is a late higher-order cognitive component associated with higher-order cognitive processes ([Donchin, 1981](#ref-donchin1981surprise); [Duncan-Johnson, 1981](#ref-duncan1981young); [McCarthy & Donchin, 1981](#ref-mccarthy1981metric); [Polich, 2007](#ref-polich2007updating), [2012](#ref-polich2012neuropsychology); [Verleger, 2020](#ref-verleger2020effects)). A number of studies have demonstrated a large effect of age on the latency of the P3 across a number of tasks with older participants displaying systematically later P3 peaks than their younger counterparts ([Friedman, 2011](#ref-friedman2012components); [Scrivano & Kieffaber, 2022](#ref-scrivano2022behavioral)). In a multiverse approach Sadus et al. ([2023](#ref-sadus2023multiverse)) tested several extraction methods with varying preprocessing steps in their ability to detect this age effect. They used three tasks, each measuring one of the executive functions proposed by Miyake et al. ([2000](#ref-miyake2000unity)). To measure the functions *updating*, *shifting* and *inhibition*, they employed an Nback, a Switching and a Flanker Task, respectively. Studying three different tasks allows us to gain insight into a larger variety of higher-order cognitive processing, improving the generalizability of our findings.

For the present work, we will restrict our analysis to extracting P3 latencies. The P3 usually has a broad and isolated structure with comparatively low influence of surrounding components ([Luck, 2014](#ref-luck2014introduction)). This makes the P3 one of the easier components for automated latency extraction approaches. After we can demonstrate proof-of-concept for P3 latency extraction, we will evaluate the algorithms ability to be extended to other ERP components.

We hope to show that a template matching approach using the grand average as a variable template can successfully extract P3 component latencies. Ideally, use of this algorithm will improve psychometric properties in comparison to prior algorithms, show high correlations with manually extracted data and present an objective and efficient way to extract ERP latencies.

# Implementation

We implemented the algorithm in MATLAB (Version 2022b) ([The Math Works, 2022](#ref-matlab2022b)). The researcher specifies the name and polarity of the component of interest as well as the measurement window. This window is used to extract the template. In order to transform the template, we use MATLABs Curve Fitting Toolbox ([The Math Works, 2022](#ref-matlab2022b)) to generate *sum of sines* functions that fit to the data with . We then add an amplitude parameter to the overall function in order to allow for scaling of the amplitude of the template and a frequency parameter to all frequency terms in the sum of sines function to allow for “squishing” or “stretching” the template along the x-Axis. The resulting variable template is described by a function with sine terms and their respective amplitude , frequency and phase .

As these transformations also change the measurement window, we chose to use the subject-level ERP as a template and keep the grand average untransformed as a signal. This reverse matching approach is only an implementation detail and does not affect any decisions made by the algorithm.

Depending on the similarity measure employed, we use different functions to find the set of optimal parameters that lead to the optimal transformation for a given subject .

## MINSQ

The MINSQ algorithm minimizes the weighted sum of squared differences between the transformed subject signal and the grand average.

This weighting vector is computed as follows

Where

denotes the measurement window, the signal strength of the th element, the maximum voltage deflection inside the measurement window and the time of the th element.

This places more emphasis on fitting the template within the measurement window specified and to places in the signal where the voltage deflection is high.

We use MATLABs *fit* function to find optimal parameters with upper and lower bounds such that . As this function may be prone to converging on local minima, we initialize 5 different start points. The solution with the best correlation between transformed template and signal that multiple start points converged on is selected.

In cases where the subject-level ERP only has signal with deflections opposite of the deflection of the component of interest, it may occur that the parameter . In these cases, we attempt to re-match the signal with an added parameter shifting the entire template up or down.

Should the algorithm again converge on a solution with , the latency value is set to NA.

## CORR

The CORR algorithm optimizes the parameters to produce the maximum correlation between the transformed subject-level signal and the grand average for values in the measurement window.

represents the vector of values of the transformed subject ERP that are in the measurement window, the vector of values of the grand average that are in the measurement window. As the correlation-coefficient is independent of translation and scaling, varying the parameter will not impact the correlation . We therefore just set and only optimize .

We use MATLABs *fminbnd* function to find the optimal transformation parameter maximizing . This function will estimate the objective function for all values inside the given bounds and converge on the global optimum. Hence, we do not need to initialize a number of different starting points here.

For both algorithms, we use the returned value of the parameter to transform the component latency specified by the researcher in the grand average to the component latency of the subject-level ERP signal .

## Review methods

Researchers can manually review all choices the algorithm has made in a custom-built user interface (see Figure 2). For both algorithms, we used the correlation between transformed template and signal as a fit-index. This can be used to only review those cases where the correlation between template and signal dips below a certain value, indicating low similarity between matched template and signal. We will investigate the additional benefits a manual review process provides over accepting the choices as-is or only automatically discarding those matches with correlations .



Figure 2: User Interface for Manual Review Process

# Method

The data used here were first published by Sadus et al. ([2023](#ref-sadus2023multiverse)) and are a subset of the data collected by Löffler et al. ([2022](#ref-loffler2022common)).

## Participants

The data is comprised of 30 young participants (18-21 years old, mean age = 19.37, SD age = 0.76) and 30 old participants (50-60 years old, mean age = 55.83, SD age = 2.87) representing the 30 youngest and 30 oldest participants from the overall study ([Löffler et al., 2022](#ref-loffler2022common)). All participants had normal or corrected to normal vision. None of the participants had neurological or mental disorders, used psychotropic drugs, wore a pacemaker or suffered from red-green color vision deficiency. The participants provided informed consent prior to participation and received 75€ or course credit for participation.

## Tasks

Participants completed a set of 3 tasks: a Flanker Task, an Nback Task and a Switching Task. These tasks each measure one of the executive functions proposed by Miyake et al. ([2000](#ref-miyake2000unity)). Löffler et al. ([2022](#ref-loffler2022common)) programmed all tasks in MATLAB ([The Math Works, 2022](#ref-matlab2022b)) using the software package Psychtoolbox (Version 3-0.13) ([Brainard & Vision, 1997](#ref-brainard1997psychophysics); [Kleiner et al., 2007](#ref-kleiner2007psychtoolbox); [Pelli & Vision, 1997](#ref-pelli1997videotoolbox)). Stimuli were presented centrally on a black background. We instructed participants to respond as quickly and accurately as possible.

### Flanker Task.

We administered a standard Arrow Flanker task ([Eriksen & Eriksen, 1974](#ref-eriksen1974effects)) to measure participants’ *inhibition* ability. A central arrow pointing either to the left or to the right is flanked by two additional arrows to each side. These flanking arrows either point in the same or in the opposite direction as the central arrow. Participants have to respond by button press in which direction the central arrow pointed, disregarding the congruent or incongruent flanking arrows. Participants completed a set of practice trials and a total of 100 congruent and 100 incongruent trials.

### Nback Task.

We administered an adapted version of the Nback task from ([Scharinger et al., 2015](#ref-scharinger2015flanker)) to measure participants’ *updating* abilities. A stream of letters is shown to the participant. In the 0-back condition, participants have to indicate by keypress whether the presented letter is equivalent to a target letter. In the 1-back condition, participants have to indicate whether the currently presented letter is the same as the letter presented one trial before or not. Löffler et al. ([2022](#ref-loffler2022common)) also had participants complete a 2-back condition. We excluded this condition from our analysis as it did not produce clear ERPs. Participants completed a set of practice trials and a total of 96 trials per condition.

### Switching Task.

We administered a Switching task to measure participants’ *shifting* ability. A stream of colored digits ranging from 1 to 9 was presented. Participants had to indicate whether the digit was greater than or less than 5 or whether the digit was odd or even depending on the color of the stimulus. A colored fixation cross just prior to stimulus presentation cued the rule participants had to follow in the upcoming trial. Participants had to either follow the same rule as in the trial before or switch to the other rule. Participants completed a set of practice trials and 192 trials each in the repeat and in the switch condition.

## Procedure

The original study consisted of three test sessions. The three tasks the present work focuses on were all administered in the first session. The second session also included EEG measurement with 3 additional tasks. The third session was used to measure intelligence and working memory capacity. No EEG measurements were taken here. In sessions including EEG measurements, participants were seated approximately 140cm away from a monitor in a sound-attenuated room.

## EEG recording and processing

EEG was recorded continuously using 32 equidistant Ag/AgCl electrodes. Additional electrooculogram (EOG) measures were taken by two electrode placed above and below the left eye to correct for ocular artifacts. All impedances were kept below 5 kΩ. The signal was recorded with a sampling rate of 1000 Hz and online-referenced to Cz. To remove artifacts, we conducted an ICA on a cloned version of the dataset down-sampled to 200 Hz and passed through an additional high-pass filter of 1 Hz. Both the original data as well as the ICA-dataset were cleaned by removing line-noise using the CleanLine function ([Mullen, 2012](#ref-mullen2012cleanline)). We also applied zvalue based bad channels detection with a critical z-value of 3.29 as recommended in the EPOS pipeline ([Rodrigues et al., 2021](#ref-rodrigues2021epos)). Channels that were removed following this procedure were interpolated and the data was re-referenced to the average across electrodes. We set as the threshold for large fluctuations and data had to be more than 5 SDs from the mean to exceed our probability threshold. Based on these settings, we automatically detected segments containing artifacts and removed those in the ICA-dataset up to a maximum of 5% of segments per iteration. ICA was conducted using the InfoMax algorithm and the resulting decomposition applied to the original dataset. ICs were labelled using the ICLabel Algorithm ([Pion-Tonachini et al., 2019](#ref-pion2019iclabel)) and removed if the IC was less than 50% likely to be brain activity. We then applied Butterworth low-pass filters with varying cut-off frequencies (8 Hz, 16 Hz, 32 Hz) and a roll-off of 12 dB/octave. Data was segmented into 1200ms long segments starting 200ms before stimulus onset. Segments containing artifacts were automatically detected and removed. As a last step, we conducted a baseline correction using the 200ms prior to stimulus onset.

## ERP analysis

ERP analyses were conducted in MATLAB (Version 2022b) ([The Math Works, 2022](#ref-matlab2022b)). We only included correct trials into analysis. We investigated the P3 at the electrode Pz, following existing literature ([Polich, 2012](#ref-polich2012neuropsychology); [Verleger, 2020](#ref-verleger2020effects)).

### Latency extraction.

To evaluate the impact of the specified measurement window, we extracted latencies three separate times using either a narrow (250-600ms), medium (200-700ms) or wide (150-900ms) measurement window. We used the peak latency approach to determine the latency of the P3 in the grand averages that is then used to recover subject-level latencies in the template matching algorithms. We applied our algorithms using both the distance-based (MINSQ) and correlation-based (CORR) similarity measures to the data and obtained transformation parameters and fit values.

To investigate the benefits of manually reviewing the decisions of the algorithm, we chose to review all matches that resulted in fit values . We then inspected the subject-level ERP with the matched template displayed over it using our interface and either accepted, rejected or manually determined the P3 latency of these ERPs. We also explored the impact of automatically excluding all those matches with fit values .

In our dataset, each of the 60 participants contributed 6 ERPs per task to the data. One ERP averaged over all trials of each of the two conditions and two more ERPs generated from an odd-even split on a trial level of that condition. These 360 ERPs each from the 3 different tasks were passed through 3 different low-pass filters and subjected to analyses with 3 separate measurement windows. We applied both the correlation-based (CORR) and distance-based (MINSQ) algorithm and either reviewed the results manually, discarded bad matches automatically or accepted the results regardless of fit. We also applied both a peak latency and 50% area latency algorithm. For the area latency algorithm, we set all values below zero to be equal to zero to combat low frequency noise ([Liesefeld, 2018](#ref-liesefeld2018estimating); [Luck, 2014](#ref-luck2014introduction)) and determined the exact latency by linear interpolation between data points. This results in 3 tasks 3 filters 3 windows algorithms = 216 different extraction pipelines.

## Validation Techniques

We investigated the impact of latency extraction method on several measures of psychometric quality. We estimated reliability by computing Spearman-Brown corrected split-half correlations of ERPs generated from an odd-even split at the trial level. We assessed the validity of our algorithm through measures of homogeneity, the effect sizes of the age effect and the correlation of latencies extracted by our algorithm with latencies extracted by an expert ERP researcher, which constitute a benchmark for proper latency extraction ([Sadus et al., 2023](#ref-sadus2023multiverse)).

To compute a methods homogeneity , we calculated its correlation with all other methods and took the mean of the Fisher-Z transformed correlation coefficients. Correlation coefficients of 1 cannot be transformed. Thus, we set all correlations to . The mean correlation with other methods indicates the extent to which a particular method reflects the total of all other measures ([Kline, 1986](#ref-kline1986handbook)).

To investigate the effect of age on P3 latencies, we ran a repeated measures ANOVA with the between factor age (young vs. old) and the within factor task (Flanker, Nback, Switching).

We quantified the extent to which an extraction method is able to replicate the benchmark of manual extraction by correlating the latencies extracted by a particular method with those extracted by an expert researcher in the same task and filter condition.

We also quantified the impact preprocessing steps and the choice of measurement window have on the reliability and validity of extraction methods. We ran a repeated measures ANOVA with the dependent variable being either the estimated reliability of latency values generated by a method or the correlation with manually extracted data. We used the between factor filter (8 Hz vs. 16 Hz vs. 32 Hz) and the within factor measurement window (narrow vs. medium vs. wide).

# Results

All data wrangling and statistical analyses were conducted using R [Version 4.1.3; R Core Team ([2022](#ref-R-base))][[1]](#footnote-57).

## Review process

We reviewed results of the CORR and MINSQ algorithms if their fit was below or if or . For the CORR algorithm, out of 9720 ERPs evaluated by the algorithm, we inspected 1045 (10.75 %). Of those ERPs, we rejected 23.35 % and accepted 64.21 % of the results despite their fit. We manually corrected the decisions in 12.44 % of cases. Automatically rejecting fits discards 2.09 % of latencies. For the MINSQ algorithm, out of 9720 ERPs evaluated by the algorithm, we inspected 1063 (10.94 %). Of those ERPs, we rejected 28.22 % of ERPs and accepted 62.65 % of the results despite their fit. We manually corrected the decisions in 9.13 % of cases. Automatically rejecting fits discards 1.43 % of latencies. Because the MINSQ algorithm may fail to find a valid solution if an amplitude parameter of fits the signal best, we discarded 7.13 % of cases. This did not occur in the CORR algorithm.

When reporting the psychometric properties of the algorithms, we will focus on those values passed through manual inspection. Values that were gained from a pipeline ending with the automatic rejection filter are reported in parenthesis. Properties of uninspected pipelines can be found in the respective tables in the appendix.

## Reliability

An overview of Spearman-Brown corrected split-half correlations split by task, measurement window and filter setting can be found in [TABLE]. Across tasks, measurement windows and filter settings the CORR algorithm had mean split-half correlations of .89 for manually reviewed latencies ( .87 for automatically reviewed latencies), the MINSQ algorithm .90 ( .86). Area latency measures presented mean split-half correlations of .92. Peak latency measures had mean split-half correlations of .81. The average split-half correlation for values extracted by an expert ERP researcher was for area latency measures and for peak latency measures ([Sadus et al., 2023](#ref-sadus2023multiverse)).

## Homogeneity

An overview of a method’s mean correlation with other methods split by task, measurement window and filter setting can be found in [TABLE]. Across tasks, measurement windows and filter settings the CORR algorithm had a homogeneity of .81 ( .80), the MINSQ algorithm .87 ( .86). The homogeneity of area latency measures was .78, and .71 for peak latency measures.

## Effect size

An overview of the effect size of the age effect detected by a particular method split by task, measurement window and filter setting can be found in [TABLE]. Across tasks, measurement windows and filter settings the CORR algorithm had a mean effect size of .18 ( .18), the MINSQ algorithm .23 ( .23). Area latency measures presented average effect sizes of .15, peak latency measures .08. The average effect size for values extracted by an expert ERP researcher was for area latency measures and for peak latency measures ([Sadus et al., 2023](#ref-sadus2023multiverse)).

## Correlation with manual rater

An overview of the correlation with latency values extracted by an expert ERP researcher ([Sadus et al., 2023](#ref-sadus2023multiverse)) split by task, measurement window and filter setting can be found in [TABLE]. Across tasks, measurement windows and filter settings the CORR algorithm had mean correlations of .89 ( .88) with manually extracted latencies, the MINSQ algorithm .91 ( .91). Area latency measures had a mean correlation of .85, peak latency measures .82.

## Influence of researcher degrees of freedom

The repeated measures ANOVA with split-half correlation of a particular method as a dependent variable, the between factor filter (8 Hz vs. 16 Hz vs. 32 Hz) and the within factor measurement window (narrow vs. medium vs. wide) showed no effect of filter, .00, for the CORR algorithm, no effect, .00, for the MINSQ algorithm no effect, .00, for area latency measures and no effect, .00, for peak latency measures. The measurement windows had an effect of .17 for the CORR algorithm, .00 for the MINSQ algorithm, .31 for area latency measures and .66 for peak latency measures. Consult [TABLE] for a full overview.

When the correlation with manually extracted latencies was used as a dependent variable, the effect of filter settings for the CORR algorithm were .00, for the MINSQ algorithm .00, .00 for area latency measures and .00 for peak latency measures. The measurement windows had an effect of .30 for the CORR algorithm, .27 for the MINSQ algorithm, .70 for area latency measures and .70 for peak latency measures. Consult [TABLE] for a full overview.

# Discussion

Our newly proposed pattern matching algorithms displayed consistently good psychometric properties and showed an improved ability to replicate human extraction behavior over previously established approaches like peak latency or area latency algorithms. Manual extraction has so far proven superior to algorithmic approaches ([Sadus et al., 2023](#ref-sadus2023multiverse)) but presents a time- and resource-intensive process. Our algorithm based on minimizing the weighted squared distance between transformed template and signal (MINSQ) correlated to .91 with manually extracted ERP latencies across tasks and preprocessing steps. This indicates that our algorithm was able to replicate manual extraction almost perfectly while presenting a more objective and efficient approach to latency extraction. Our algorithms were also more robust to the impact of different low-pass filters and measurement windows than previous algorithms. Application of our algorithm would increase both replicability and scalability as well as significantly reduce the time and resources researchers need to spend on latency extraction.

Because previous algorithms have been proposed, a key comparison was evaluate the effectiveness of our algorithm against already established, potentially simpler algorithms.

Regarding the reliability of extracted latencies across tasks and preprocessing steps, our algorithms did not prove superior to the area latency approach. Both the MINSQ and CORR approaches led to slightly lower Spearman-Brown corrected split-half correlations than in area latency approaches. However, the differences in reliability are quite small and only carry low practical implications. If the researcher uses latent-variable approaches in their analysis of ERP latencies, for example, slightly lower reliabilites don’t influence the quality of results much [REF].

Latency values extracted by the MINSQ algorithm proved to have the highest average correlation with all other extraction methods across tasks and preprocessing steps ( .87). This indicates that this approach best reflects the total of all other measures. The CORR algorithm also proved superior to previously established extraction methods.

Sadus et al. ([2023](#ref-sadus2023multiverse)) showed that manually extracting latency values is the best approach to ensure good psychometric properties and high power to detect experimental effects. The ability of an algorithm to extract latency values correlating highly with those extracted by an expert ERP-researcher was therefore of high importance to us. Again, our algorithms proved to have a superior ability to replicate human behavior compared to previous approaches. The MINSQ algorithm, after manual inspection, had a mean correlation of .91 with manually extracted latencies across tasks and preprocessing steps. Considering the reliabilites of the two algorithms, .90 and , this correlation exceeds the theoretically maximal correlation between their true-score values [REF]. This indicates that some amount of their measurement error covaries. We suspect that this is due to the high conceptual similarity between template matching and manual extraction. Errors in the generation of the template or errors related to the matching procedure may be similar for human researchers and the algorithm.

The CORR algorithm also outperformed previously established approaches in the ability to replicate human behavior, correlating very highly with manually extracted latencies ( .89). Area latency measures also correlate highly with manually extracted data ( .85) but failed to match the performance of our new algorithms.

Importantly, even the fully automated pipelines of our algorithms that rejected those matches with fits led to mean correlations of .91 for the MINSQ algorithm and .88 for the CORR algorithm. This indicates that our algorithm may be applied fully autonomously, eliminating the need for human intervention and increasing replicability and efficiency.

Ideally, the quality of an algorithm extracting ERP latencies is largely independent of choices made by the researcher during preprocessing. To evaluate this, we conducted ANOVAs with either the split-half correlation as an estimate for reliability or the correlation of a particular method with manually extracted latencies as an estimate of validity as the dependent variable. The filter setting (8 Hz vs. 16 Hz vs. 32 Hz) and the measurement window (narrow vs. medium vs. wide) were entered as independent variables. The filter setting used did not have an impact on any of the two dependent measures for any algorithm. The measurement window impacted the reliability of peak latency and area latency the most, only the reliability of the MINSQ algorithm was not impacted by the choice of measurement window. The impact of the measurement window on the correlation with manually extracted latencies was again greatest for peak latency and area latency and much lower for the CORR and MINSQ algorithms. Overall, our new algorithms proved to be more robust to influences of the measurement window. This is likely due to the lower conceptual influence of the measurement window in our algorithms. While the measurement window directly impacts the values of each subject-level ERP entered into analysis in prior algorithms, it only determines the size of the template used in our algorithms.

Any algorithmic approach to ERP latency extraction will be more objective than manually extracting ERP latencies. So we cannot crown any particular algorithm as more or less objective. The completely autonomous versions of peak latency, area latency or our algorithm with automatic rejection of bad fits are all equally objective. One strength of our approach is the ability for the researcher to inspect a subset of the ERPs based on the fit statistic of the matching procedure. This does introduce some subjectivity into the data.

However, this ability of our algorithms to generate a fit statistic indicating the degree of certainty with which the match was made is a great strength of our new algorithm. Depending on the size of their data and the degree of certainty to which researchers want to manually inspect their data, one may choose any cut-off value for the fit statistic and inspect none, a subset or all of the ERPs and the choices made by the algorithm by hand. This feature is not present in any of the previous algorithms.

## Comparing MINSQ and CORR

We chose to test out two different approaches to quantifying the degree of similarity between template and signal. One minimizing the weighted squared difference and one maximizing the correlation between the two. Both showed improvements over previous algorithms and we can recommend that both approaches be studied further. We did observe some differences between the two approaches in both procedural factors as well as outcome measures.

Procedurally, the largest difference between the two approaches is the optimization algorithm underlying them. Due to the invariance of the correlation of two vectors to scaling in amplitude of one vector, we can reduce the number of free parameters optimized during the CORR approach to one. This allows us to use a more exhaustive optimization algorithm that will find the global optimum in some bounded parameter space without the possibility of converging on a local optimum. This is not the case for the multivariate optimization function needed for the MINSQ approach. Here, we initialize the optimization process at several different starting points and check for convergence on a common solution indicating that this solution represents the true global optimum. This is not ideal and could be improved in the future by implementing a more suitable optimization algorithm or improving on the one currently used.

The MINSQ algorithm may also converge on solutions where if the subject level signal is largely of a polarity opposite to that of the component of interest [see Figure 3]. Although we did extend the variability of the template by a parameter vertically shifting the template to account for these cases, sometimes even the extended version will converge on solutions with non-sensible parameter values. This leads to missing values and unreliable fit statistics in those cases. In our data, this happened for 7.13 % of ERP signals. A large proportion of these cases may be considered *unidentifiable* even by an expert researcher due to particularly low signal-to-noise ratios. However, few cases where the component can be identified by a human researcher or the CORR algorithm may be classified as missing by the MINSQ algorithm. We will implement additional measures aiming to reduce the number of cases where the MINSQ algorithm fails to converge on a valid solution in future work.



Figure 3: Subject ERP with no positive-going signal

This leads to the difference in the number of cases classified as missing by the MINSQ and CORR algorithms. While 2.52 % of cases were set to NA after manual inspection of the CORR algorithm (2.09 % after automatic inspection), 7.42 % of cases were set to NA in the MINSQ algorithm following inspection (8.56 % after automatic inspection). This tradeoff between better properties of the MINSQ algorithm accompanied by more missing values must be taken into account when selecting which algorithm to use. Depending on the number of participants available and the means of analysis, missing values may be detrimental, leading to the CORR algorithm being the preferable choice.

The weighting vector used in the MINSQ algorithm represents another difference between the two algorithms. We used it to reflect the increased emphasis a human researcher places on those parts of the signal with the highest amplitude and signal appearing in the measurement window where the component of interest is expected to occur. The particular shape of the weighting function is somewhat arbitrary, but general aspects were chosen to reflect a few key considerations. For example, the maximum-normalization conducted before weights are calculated ensures that the weighting function is scale-invariant. Furthermore, we added larger weights to values inside the measurement window without completely discarding the impact of values outside the measurement window. We also chose to square the normalized amplitude in order to reflect a non-linear relationship between amplitude and importance. The exact shape of this weighting function may be argued and optimized further.

Regarding outcome measures, the MINSQ algorithm dominates the CORR algorithm in almost all indices inspected. It has better reliability, homogeneity, validity and is more robust to the impact of measurement windows. This provides evidence towards the argument that the MINSQ algorithm presents the better choice if one is limited to the application of just one algorithm.

## The impact of manual inspection

In order to choose a cut-off value for the fit statistic we tested different cut-off values and checked whether a large enough proportion of them proved problematic enough to merit manual inspection. We set as our cut-off because more conservative cut-offs led to us inspecting a larger proportion of matches with clearly correct results. Considering the size of our data and the number of ERPs we applied the algorithm to, we wanted to test our algorithm in its ability to allow an efficient extraction of latencies even in the face of a large dataset. We inspected around 10.75 % of ERPs of the CORR algorithm and 10.94 % of ERPs in the MINSQ algorithm. Depending on how liberal or conservative the inspection should be conducted, one can adjust the cut-off value to increase or decrease the percentage of ERPs to be manually inspected.

This additional effort of manual inspection led to improved qualities of the extracted latencies over just automatically discarding fits with very bad fit statistics. Mean reliability and homogeneity improved and the values had slightly higher correlations with manually extracted latencies. However, the automatic rejection filter of was still able to extract latency values better than previously established algorithms and showed mean correlations with an expert researcher of .88 for the CORR and .91 for the MINSQ algorithm.

Quantifying the certainty with which the template matching procedure chose a particular solution sets our algorithm apart from previous approaches. This enables the researcher to choose cut-off values for manual inspection and automatic rejection based on their particular needs in the current study. While more conservative inspection and rejection criteria will most likely improve the qualities of the extraction method, it also increases the time spent on inspection or the number of unidentifiable subject-level ERPs. This degree of control, especially using an objective criterion, is not available to researchers using other approaches.

Recently, Luck et al. ([2021](#ref-luck2021standardized)) introduced the Standardized Measurement Error (SME) quantifying the quality of an individual’s ERP data. This can be used to prune the dataset automatically, increasing the reliability of the latency values generated ([Wascher et al., 2022](#ref-wascher2022mental)). Future work will investigate the benefit of incorporating the SME into our review process.

## Limitations

Our template matching algorithm is limited by the type of transformation we employ to introduce variability mapping individual differences. For example, we chose not to implement a parameter shifting the entire template along the x-axis. Thus, in our algorithm the only way latency can be shifted is by also scaling the entire component. Later peaks thus necessitate broader components. This could be changed by introducing a parameter shifting the template without scaling it. However, this would also move the amplitude at 0 ms to some other time-point. As the origin is of special importance in ERP research, we decided against this shifting parameter. It is the only fix-point resulting from the averaging and baselining procedures. Thus, we chose not to disturb this property. Future work may investigate the impact the introduction of this additional parameter in template transformations has on the template matching algorithm. We also limited the algorithm to linear transformations of the template but could easily extend it to include non-linear scaling as well. Non-linear scaling would enable the template transformations to capture the effect of some participants not displaying speed differences in early components (low scaling), but showing slow late components (higher scaling).

We only inspected one cut-off value for manual inspection and one for automatic rejection. These values were based on our experience in working with the algorithm, but this only provides limited insight into the impact of the cut-off value. Choosing a more conservative automatic rejection criterion may improve reliability and validity even further but come at a cost of a larger amount of missing values.

The generalizability of our findings is limited by the data we analyzed here. We inspected a limited sample of participants, narrow range of tasks and only one ERP component. Depending on the component of interest, the effectiveness of different algorithms can vary ([Kiesel et al., 2008](#ref-kiesel2008measurement); [Wascher et al., 2022](#ref-wascher2022mental)). We suspect that the effectiveness of all algorithms will decline when attempting to extract earlier components. The P3 is a broad, high-amplitude and isolated component. This renders it ideal for algorithmic approaches, as the influence of surrounding components is comparatively low and the measurement window quite easily specified. Especially area latency approaches should diminish in quality due to the less isolated component structure of earlier components ([Luck, 2014](#ref-luck2014introduction)). We therefore expect that the benefits of our new algorithm relative to established algorithms will increase in earlier components.

## Future research

Future research should focus on applying template matching algorithms to earlier components. We also suggest simulating data to be able to quantify the algorithm’s ability to recover the true latency of a component. This present work serves largely as a proof-of-concept. The algorithms presented here have yet to prove themselves in a larger variety of tasks, samples and for different ERP components.

We also suggest improving the optimization processes used during our algorithms. The function used to implement the optimization of the MINSQ does not consistently converge on the global optimum, which we compensated for by initializing five different starting points and testing the solutions for convergence. This could be improved upon further.

Currently, the algorithms aim to identify this global optimum representing the absolute best similarity between transformed template and signal. It may be advantageous to use a linear combination of the best percentile of transformations as the solution of the optimization process ([Brunelli, 2009](#ref-brunelli2009template); [Brunelli & Poggiot, 1997](#ref-brunelli1997template)). Brunelli ([2009](#ref-brunelli2009template)) raised this issue in the context of multiclass pattern recognition. Correlation filters tend to result in broad peaks of optimality. We currently just choose the absolute peak and the algorithm returns the corresponding transformation parameters. Choosing the highest point in that peak is influenced by noise in the same manner as picking peaks of ERP components with peak latency algorithms. Future research should investigate using a linear transformation like a weighted average.

Aside from improvements in the implementation of the algorithm and extensions of the algorithms to earlier components, we will also improve the user interface employed for manual inspection. Currently, the interface displays the matched template and informs the researcher about the latency and fit statistic this match would result in. We also display the choices a peak latency and an area latency algorithm would have made. The researcher can then either accept the matched result, choose a result of the older algorithms, manually specify the component latency or reject the ERP overall due to poor identifiability. We will aim to improve this by adding a slider controlling the transformation parameters, allowing the researcher to manually match the template to the subject-level ERP. The functionality of manual latency specification will also be improved by integrating already existing software like the Measurement Tool provided by ERPLAB ([Lopez-Calderon & Luck, 2014](#ref-lopez2014erplab)).

The particular cut-off values we chose for manual inspection or automatic rejection of the template matching solution allowed us to demonstrate both the algorithm’s ability to extract ERP latencies completely automatically and the improvements gained from manually inspecting a subset of the choices made. However, we did not quantify how different cut-off values would impact the number of ERPs inspected or rejected and the resulting quality of the extraction method. We will investigate this in further research, quantifying the impact of different cut-off values in order to gain insight into which cut-off values we can recommend depending on the context in which our algorithms are applied.

# Conclusion

This work provides proof-of-concept showing that template matching algorithms using the grand average as a template can be feasibly used to extract P3 latencies. Latencies extracted by our algorithm correlate highly with values extracted by an expert human researcher across tasks and preprocessing steps. Our algorithm is superior to previous algorithms like peak latency and area latency regarding the correlation with manually extracted latencies, homogeneity and robustness to the influence of different measurement windows. A main benefit of our approach is the ability to quantify the algorithm’s confidence in a particular solution via a fit statistic. This allows researchers to inspect only the subset of ERPs with the worst fits and thus correct potential measurement error of the algorithm in a time-efficient manner. It also allows specification of a cut-off value for automatically rejecting template matches with bad fits, eliminating the need for human intervention. This fully automatic approach also displays qualities superior to previous algorithm. When comparing our two similarity measures, the MINSQ algorithm displays better qualities than the CORR algorithm. However, it also results in a higher number of missing values. We will aim to improve the implementation of our algorithms and test their ability to extract earlier ERP components. Overall, the results obtained here leave us optimistic regarding the applicability of this template matching approach. It provides a more objective and efficient way to extract ERP latencies while maintaining consistently good psychometric quality and almost perfectly replicating decisions made by an expert human researcher.

# References

Anderson, J. R., Zhang, Q., Borst, J. P., & Walsh, M. M. (2016). The discovery of processing stages: Extension of Sternberg’s method. *Psychological Review*, *123*(5), 481.

Aust, F., & Barth, M. (2022). *papaja: Prepare reproducible APA journal articles with R Markdown*. <https://github.com/crsh/papaja>

Borst, J. P., & Anderson, J. R. (2015). The discovery of processing stages: Analyzing EEG data with hidden semi-Markov models. *NeuroImage*, *108*, 60–73.

Brainard, D. H., & Vision, S. (1997). The psychophysics toolbox. *Spatial Vision*, *10*(4), 433–436.

Briechle, K., & Hanebeck, U. D. (2001). Template matching using fast normalized cross correlation. *Optical Pattern Recognition XII*, *4387*, 95–102.

Brunelli, R. (2009). *Template matching techniques in computer vision: Theory and practice*. John Wiley & Sons.

Brunelli, R., & Poggiot, T. (1997). Template matching: Matched spatial filters and beyond. *Pattern Recognition*, *30*(5), 751–768.

Clayson, P. E., Baldwin, S. A., & Larson, M. J. (2013). How does noise affect amplitude and latency measurement of event-related potentials (ERPs)? A methodological critique and simulation study. *Psychophysiology*, *50*(2), 174–186. <https://doi.org/10.1111/psyp.12001>

Cowley, B. U., Korpela, J., & Torniainen, J. (2017). Computational testing for automated preprocessing: A Matlab toolbox to enable large scale electroencephalography data processing. *PeerJ Computer Science*, *3*, e108. <https://doi.org/10.7717/peerj-cs.108>

Donchin, E. (1981). Surprise!… surprise? *Psychophysiology*, *18*(5), 493–513.

Donchin, E., & Heffley, E. F. (1978). *Multivariate analysis of event-related potential data: A tutorial review*.

Duncan-Johnson, C. C. (1981). Young psychophysiologist award address, 1980: P300 latency: A new metric of information processing. *Psychophysiology*, *18*(3), 207–215.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*(1), 143–149. <https://doi.org/10.3758/BF03203267>

Friedman, D. (2011). The components of aging. In *The oxford handbook of event-related potential components*. Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780195374148.013.0243>

Goshtasby, A., Gage, S. H., & Bartholic, J. F. (1984). A two-stage cross correlation approach to template matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *3*, 374–378.

Kiesel, A., Miller, J., Jolicœur, P., & Brisson, B. (2008). Measurement of ERP latency differences: A comparison of single-participant and jackknife-based scoring methods. *Psychophysiology*, *45*(2), 250–274.

Kleiner, M., Brainard, D., & Pelli, D. (2007). *What’s new in psychtoolbox-3?*

Kline, P. (1986). *A handbook of test construction: Introduction to psychometric design. New York: Methuen*. Inc.

Lenth, R. V. (2023). *Emmeans: Estimated marginal means, aka least-squares means*. <https://CRAN.R-project.org/package=emmeans>

Lewis, J. P. (1995). Fast template matching. *Vision Interface*, *95*, 15–19.

Li, Y., Ma, Z., Lu, W., & Li, Y. (2006). Automatic removal of the eye blink artifact from EEG using an ICA-based template matching approach. *Physiological Measurement*, *27*(4), 425.

Liesefeld, H. R. (2018). Estimating the timing of cognitive operations with MEG/EEG latency measures: A primer, a brief tutorial, and an implementation of various methods. *Frontiers in Neuroscience*, *12*, 765.

Löffler, C., Frischkorn, G. T., Hagemann, D., Sadus, K., & Schubert, A.-L. (2022). *The common factor of executive functions measures nothing but speed of information uptake*.

Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: An open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience*, *8*, 213.

Luck, S. J. (2005). Ten simple rules for designing and interpreting ERP experiments. *Event-Related Potentials: A Methods Handbook*, *4*.

Luck, S. J. (2014). *An introduction to the event-related potential technique*. MIT press.

Luck, S. J., Stewart, A. X., Simmons, A. M., & Rhemtulla, M. (2021). Standardized measurement error: A universal metric of data quality for averaged event-related potentials. *Psychophysiology*, *58*(6), e13793. <https://doi.org/10.1111/psyp.13793>

Mahalakshmi, T., Muthaiah, R., & Swaminathan, P. (2012). Image processing. *Research Journal of Applied Sciences, Engineering and Technology*, *4*(24), 5469–5473.

McCarthy, G., & Donchin, E. (1981). A metric for thought: A comparison of P300 latency and reaction time. *Science (New York, N.Y.)*, *211*(4477), 77–80.

Meyer, D. E., Osman, A. M., Irwin, D. E., & Yantis, S. (1988). Modern mental chronometry. *Biological Psychology*, *26*(1), 3–67. <https://doi.org/10.1016/0301-0511(88)90013-0>

Miller, J., Patterson, T., & Ulrich, R. (1998). Jackknife-based method for measuring LRP onset latency differences. *Psychophysiology*, *35*(1), 99–115. <https://doi.org/10.1111/1469-8986.3510099>

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex “frontal lobe” tasks: A latent variable analysis. *Cognitive Psychology*, *41*(1), 49–100.

Mullen, T. (2012). CleanLine EEGLAB plugin. *San Diego, CA: Neuroimaging Informatics Toolsand Resources Clearinghouse (NITRC)*.

Pelli, D. G., & Vision, S. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437–442.

Pion-Tonachini, L., Kreutz-Delgado, K., & Makeig, S. (2019). ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage*, *198*, 181–197.

Polich, J. (2007). Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology*, *118*(10), 2128–2148.

Polich, J. (2012). Neuropsychology of P300. *The Oxford Handbook of Event-Related Potential Components*, 159–188.

Posner, M. I. (2005). Timing the Brain: Mental Chronometry as a Tool in Neuroscience. *PLOS Biology*, *3*(2), e51. <https://doi.org/10.1371/journal.pbio.0030051>

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>

Rodrigues, J., Weiß, M., Hewig, J., & Allen, J. J. (2021). EPOS: EEG processing open-source scripts. *Frontiers in Neuroscience*, *15*, 660449.

Sadus, K., Schubert, A.-L., Löffler, C., & Hagemann, D. (2023). *A multiverse study for extracting differences in P3 latencies between young and old adults*.

Scharinger, C., Soutschek, A., Schubert, T., & Gerjets, P. (2015). When flanker meets the n-back: What EEG and pupil dilation data reveal about the interplay between the two central-executive working memory functions inhibition and updating. *Psychophysiology*, *52*(10), 1293–1304. <https://doi.org/10.1111/psyp.12500>

Schubert, A.-L., Löffler, C., Hagemann, D., & Sadus, K. (2023). How robust is the relationship between neural processing speed and cognitive abilities? *Psychophysiology*, *60*(2), e14165.

Scrivano, R. M., & Kieffaber, P. D. (2022). Behavioral and electrophysiological correlates of Simon and flanker conflict interference in younger and older adults. *Aging, Neuropsychology, and Cognition*, *29*(2), 318–348.

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2023). *Afex: Analysis of factorial experiments*. <https://CRAN.R-project.org/package=afex>

Smulders, F. T. Y. (2010). Simplifying jackknifing of ERPs and getting more out of it: Retrieving estimates of participants’ latencies. *Psychophysiology*, *47*(2), 387–392. <https://doi.org/10.1111/j.1469-8986.2009.00934.x>

The Math Works, Inc. (2022). *MATLAB version: 9.13.0 (r2022b)*. The MathWorks Inc. <https://www.mathworks.com>

Ulrich, R., & Miller, J. (2001). Using the jackknife-based scoring method for measuring LRP onset effects in factorial designs. *Psychophysiology*, *38*(5), 816–827.

Verleger, R. (2020). Effects of relevance and response frequency on P3b amplitudes: Review of findings and comparison of hypotheses about the process reflected by P3b. *Psychophysiology*, *57*(7), e13542.

Wascher, E., Sharifian, F., Gutberlet, M., Schneider, D., Getzmann, S., & Arnau, S. (2022). Mental chronometry in big noisy data. *PLOS ONE*, *17*(6), e0268916. <https://doi.org/10.1371/journal.pone.0268916>

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. <https://doi.org/10.21105/joss.01686>

William, F., Aygun, R., & Zhu, F. (2020). ERP template matching for EEG single trial classification. *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2876–2883.

Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Chapman; Hall/CRC. <https://yihui.org/knitr/>

Xie, Y., Allaire, J. J., & Grolemund, G. (2018). *R markdown: The definitive guide*. Chapman; Hall/CRC. <https://bookdown.org/yihui/rmarkdown>

Xie, Y., Dervieux, C., & Riederer, E. (2020). *R markdown cookbook*. Chapman; Hall/CRC. <https://bookdown.org/yihui/rmarkdown-cookbook>

# Appendix

# Algorithm Quality

Table 1: Reliability of different algorithms - Flanker Task

| filter | window | corr | | | minsq | | | autoarea | autopeak |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| none | auto | manual | none | auto | manual | none | none |
| 8 | narrow | 0.96 | 0.95 | 0.95 | 0.93 | 0.93 | 0.93 | 0.96 | 0.91 |
| medium | 0.97 | 0.97 | 0.97 | 0.87 | 0.87 | 0.95 | 0.95 | 0.84 |
| wide | 0.88 | 0.88 | 0.96 | 0.95 | 0.97 | 0.94 | 0.94 | 0.82 |
| 16 | narrow | 0.95 | 0.94 | 0.94 | 0.90 | 0.90 | 0.92 | 0.96 | 0.85 |
| medium | 0.97 | 0.97 | 0.97 | 0.80 | 0.80 | 0.86 | 0.95 | 0.81 |
| wide | 0.88 | 0.87 | 0.96 | 0.93 | 0.95 | 0.97 | 0.94 | 0.81 |
| 32 | narrow | 0.96 | 0.96 | 0.95 | 0.92 | 0.92 | 0.96 | 0.96 | 0.84 |
| medium | 0.95 | 0.95 | 0.97 | 0.92 | 0.92 | 0.97 | 0.95 | 0.80 |
| wide | 0.94 | 0.96 | 0.96 | 0.91 | 0.91 | 0.97 | 0.94 | 0.63 |
| *Note.* My personal note on this table. | | | | | | | | | |

Table 2: Reliability of different algorithms - Nback Task

| filter | window | corr | | | minsq | | | autoarea | autopeak |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| none | auto | manual | none | auto | manual | none | none |
| 8 | narrow | 0.81 | 0.83 | 0.82 | 0.95 | 0.95 | 0.95 | 0.91 | 0.80 |
| medium | 0.82 | 0.88 | 0.83 | 0.75 | 0.76 | 0.89 | 0.93 | 0.79 |
| wide | 0.65 | 0.69 | 0.90 | 0.90 | 0.93 | 0.93 | 0.88 | 0.64 |
| 16 | narrow | 0.75 | 0.75 | 0.76 | 0.80 | 0.84 | 0.87 | 0.88 | 0.76 |
| medium | 0.78 | 0.78 | 0.81 | 0.86 | 0.86 | 0.93 | 0.92 | 0.76 |
| wide | 0.66 | 0.74 | 0.75 | 0.87 | 0.88 | 0.91 | 0.86 | 0.69 |
| 32 | narrow | 0.75 | 0.77 | 0.85 | 0.80 | 0.80 | 0.78 | 0.84 | 0.76 |
| medium | 0.64 | 0.71 | 0.71 | 0.69 | 0.69 | 0.88 | 0.92 | 0.80 |
| wide | 0.54 | 0.61 | 0.76 | 0.82 | 0.85 | 0.93 | 0.88 | 0.68 |
| *Note.* My personal note on this table. | | | | | | | | | |

Table 3: Reliability of different algorithms - Switching Task

| filter | window | corr | | | minsq | | | autoarea | autopeak |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| none | auto | manual | none | auto | manual | none | none |
| 8 | narrow | 0.70 | 0.82 | 0.79 | 0.65 | 0.71 | 0.84 | 0.94 | 0.90 |
| medium | 0.76 | 0.92 | 0.93 | 0.80 | 0.84 | 0.90 | 0.91 | 0.89 |
| wide | 0.65 | 0.69 | 0.66 | 0.71 | 0.72 | 0.84 | 0.86 | 0.81 |
| 16 | narrow | 0.61 | 0.77 | 0.80 | 0.40 | 0.70 | 0.86 | 0.95 | 0.84 |
| medium | 0.89 | 0.94 | 0.93 | 0.86 | 0.89 | 0.86 | 0.91 | 0.86 |
| wide | 0.65 | 0.67 | 0.66 | 0.81 | 0.81 | 0.81 | 0.88 | 0.82 |
| 32 | narrow | 0.76 | 0.92 | 0.87 | 0.63 | 0.82 | 0.80 | 0.89 | 0.86 |
| medium | 0.85 | 0.91 | 0.90 | 0.78 | 0.82 | 0.77 | 0.92 | 0.86 |
| wide | 0.68 | 0.73 | 0.72 | 0.65 | 0.64 | 0.63 | 0.86 | 0.83 |
| *Note.* My personal note on this table. | | | | | | | | | |

# Correlation Matrices

1. We, furthermore, used the R-packages *afex* (Version 1.3.0; [Singmann et al., 2023](#ref-R-afex)), *emmeans* (Version 1.8.9; [Lenth, 2023](#ref-R-emmeans)), *knitr* (Version 1.45; [Xie, 2015](#ref-R-knitr)), *papaja* (Version 0.1.1.9001; [Aust & Barth, 2022](#ref-R-papaja)), *rmarkdown* (Version 2.25; [Xie et al., 2018](#ref-R-rmarkdown_a), [2020](#ref-R-rmarkdown_b)), and *tidyverse* (Version 2.0.0; [Wickham et al., 2019](#ref-R-tidyverse)). [↑](#footnote-ref-57)