Psychophysiology, 45 (2008), 250–274. Blackwell Publishing Inc. Printed in the USA.
Copyright © 2007 Society for Psychophysiological Research
DOI: 10.1111/j.1469-8986.2007.00618.x

Measurement of ERP latency differences: A comparison of single-participant and jackknife-based scoring methods

ANDREA KIESEL,[a] JEFF MILLER,[b] PIERRE JOLICŒUR,[c] AND BENOIT BRISSON[c]

[a]Department of Psychology, University of Würzburg, Würzburg, Germany
[b]Department of Psychology, University of Otago, Dunedin, New Zealand
[c]Département de Psychologie, Université de Montreal, Succursale Centre-ville, Montreal, Quebec, Canada

Abstract

We used computer simulations to evaluate different procedures for measuring changes in the onset latency of a representative range of event-related components (the auditory and visual N1, P3, N2pc, and the frequency-related P3 difference wave). These procedures included several techniques to determine onset latencies combined with approaches using both single-participant average waveforms and jackknife-subsample average waveforms. In general, the jackknife-based approach combined with the relative criterion technique or combined with the fractional area technique (J.C. Hansen & S.A. Hillyard, 1980; S.J. Luck, 2005) provided the most accurate method and the greatest statistical power, with no inflation of Type I error rate.

Descriptors: ERP latency, Jackknife, N1, P3, N2pc, Frequency-related P3

Research using event-related potentials (ERPs) often focuses on differences in amplitudes as well as differences in latencies of ERP components. However, for both measurements there are several scoring methods, and quite often researchers have to decide rather arbitrarily which method might be most appropriate. In this article we compare several methods for determining ERP latency differences. To evaluate the methods we ran computer simulations based on data of five ERP components: the visual N1, the auditory N1, the P3 (hereafter used to mean the P3b), the N2pc, and the frequency-related P3 component (infrequent minus frequent difference wave).

These five components were chosen because they are very different and representative of a broad range of components, as is apparent from the following brief review of these components. The N1 components (visual, auditory) are relatively early components, strongly influenced by physical properties of the stimulus. They are characterized by a clear onset and a sharply increasing amplitude, but clearly reflect different underlying neural generators. In contrast, the P3 is a late component that is relatively insensitive to the physical properties of the stimulus (with the exception of tone intensity; see Covington & Polich, 1996), but is influenced by probabilities, expectations, and resource allocation (see Johnson, 1986). Quite often the P3 does not show a clear onset, and its peak latency is difficult to determine because the component has a wide temporal extension without a sharp peak. The N2pc and frequency-related P3, for their part, are measured from difference waves that are obtained by subtracting ERP waveforms computed at different electrode sites or at same electrode sites but in different experimental conditions. The N2pc is an index of the allocation of visual–spatial attention and is isolated by subtracting the ERP at posterior electrode sites ipsilateral to an attended item from the ERP at the corresponding contralateral electrode site, whereas the frequency-related P3 is isolated by subtracting the ERP elicited by frequent targets from the ERP elicited by infrequent targets.

For all components, researchers are often interested in estimating the latency differences across different experimental conditions. The most straightforward procedure for determining latency differences, and indeed the one most often used, is to compare peak latencies (e.g., Jemel et al., 2003; Leuthold & Sommer, 1998; Luck, 2005). To test for significance, parametric tests are applied; that is, peak latencies are measured for each condition within single-participant average waveforms, and these latency data are subjected to *t* tests or ANOVAs.

However, when asking whether ERP latencies differ between conditions, it is questionable whether the most efficient and reliable procedure is to determine peak latencies for single participants. Considering data of single participants might be

This research was supported by a grant of the G.A. Lienert Foundation to Andrea Kiesel and by a grant from The Marsden Fund administered by the Royal Society of New Zealand to Jeff Miller and by support from the Natural Sciences and Engineering Research Council of Canada and the Canada Research Chairs Program awarded to Pierre Jolicœur. We thank Albrecht Sebald, Annika Wagener, Ines Jentzsch and one anonymous reviewer for helpful comments on an earlier version of the article.

Address reprint requests to: Andrea Kiesel, Department of Psychology, University of Würzburg, Röntgenring 11, 97070 Würzburg, Germany. E-mail: kiesel@psychologie.uni-wuerzburg.de

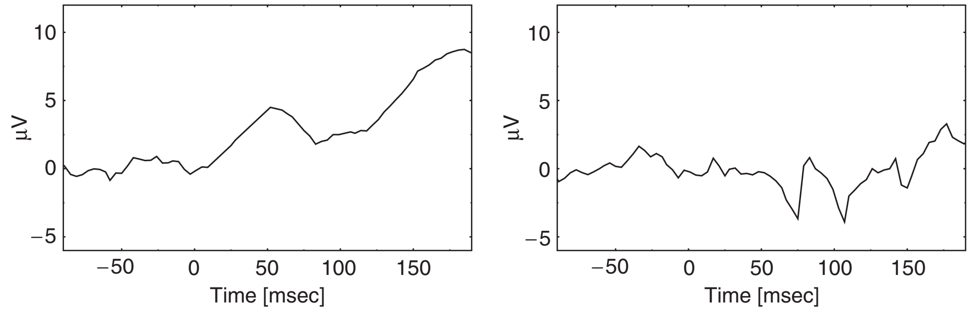

Figure 1. Hypothetical N1 data for single participants. On the left side, the minimum in the time window of the N1 is not negative. On the right side, there are two minima in the time window of the N1.

problematic for several reasons: First, estimates of peak latencies separately for each participant can be afflicted with relatively large error, because EEG signals often have a low signal-to-noise ratio. Second, the determination of a peak latency for each participant can be quite difficult if the waveform does not have the canonical shape. Consider the examples depicted in Figure 1. Suppose one wanted to estimate the latency of a hypothetical N1 component with a negative peak somewhere between 50 and 150 ms. The hypothetical data pattern on the left side does not show any negative peak in the time window of the N1. On the right side, the hypothetical data pattern reveals two peaks in the N1 time window (analogous examples could be easily constructed for the other components as well). For such cases, it is not obvious how one should determine peak latencies: In the case on the left, for example, should one just consider the time point of the minimum of the curve as N1 latency, despite the fact that the amplitude of this minimum is not negative? In the case on the right, should one take the latency of the first or of the second peak? Or should one omit the data of these participants from the analysis?

Another problem with peak latency is that it is questionable whether this is the best technique to use in judging whether ERPs differ in time. Peak latency happens long after the onset of the ERP component, so it might not be particularly sensitive to changes in the starting time of this component. To illustrate this problem, Figure 2 depicts two examples of how N1 curves might differ in an experimental and a control condition. On the left side, the whole ERP curve in the experimental condition is shifted along the time axis. Thus, the N1 component occurs later than in the control condition. In this case, the peak latency difference does seem to reflect the actual temporal shift. On the right side, however, the N1 *onsets* do appear to differ between the experimental and control conditions, yet the N1 peak latencies do not differ due to changes in the shape of the components. In the latter case, then, the peak latency technique would clearly not be appropriate to reveal onset latency differences (similar examples can be easily constructed for all other components). A second and more technical problem with peak latencies is that interpolation cannot be used with them. Peak latency estimates depend on the sampling rate of the EEG measurement and are always integer multiples of the discrete time steps at which EEG was recorded. The graininess of the latency measurement limits precision. Third, especially for a broad component like the P3, the peak, in a particular time window, could end up at the edge of the window if the window includes a rising or falling edge of the component, raising doubts as to the validity of that particular measurement (see Luck, 2005, for additional difficulties with the peak latency technique).

In this article, we evaluated several different procedures for determining whether N1 (visual, auditory), P3, N2pc, and frequency-related P3 latencies differ across two conditions. We compared the accuracy of the peak latency technique against several other scoring techniques that could be used to determine and compare ERP onset latencies. All of these scoring techniques were combined with two statistical approaches: In addition to considering scoring the data of single participants, we examined the scoring of averaged data sets by applying the jackknife approach (Miller, Patterson, & Ulrich, 1998; Ulrich & Miller, 2001). Our goal was not to establish that a single scoring procedure is the best for all of the components or even for all situations with a single component, but rather to get a basic picture of the statistical properties of these different scoring

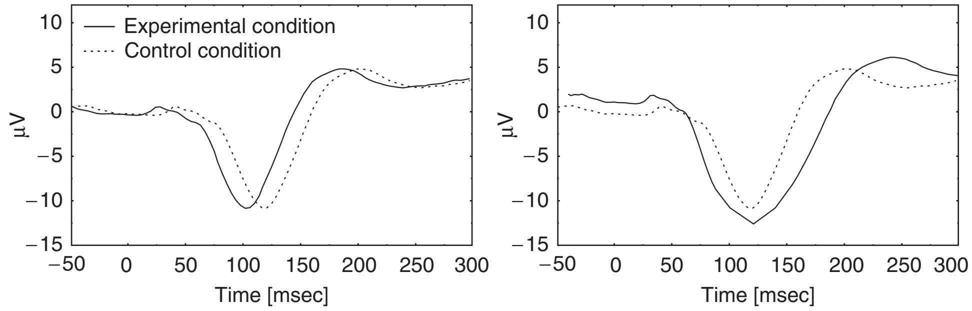

Figure 2. Hypothetical N1 latency differences. On the left side, experimental and control conditions differ because the whole N1 curve is shifted in time. On the right side, N1 onsets in the experimental and control conditions differ, but peak latencies are the same.

procedures that would enable researchers to make more informed choices about the optimal method for analyzing a specific data set.

The Jackknife Approach

There are good reasons to suspect that the jackknife approach (e.g., Efron, 1981; Jackson, 1986; Miller, 1974; Mosteller & Tukey, 1977) may provide more accurate estimates of latency differences than the approach of scoring of single-participant waveforms. With the jackknife approach, latencies are scored for each of n grand average waveforms, with each of the grand average waveforms computed from a subsample of $n-1$ of the n individual participants (i.e., each participant is omitted from one of the subsample grand averages). To test for the statistical significance of latency differences, the observed latency difference values for the n subsamples, D_{-i}, for $i = 1, 2, \ldots, n$, are computed, where D_{-i} is the latency difference obtained for the subsample of participants including all participants except participant i. The values D_{-i} are then submitted to a conventional t test or an analysis of variance (ANOVA), but before testing for significance, the t values or the F values must be adjusted according to

$$\begin{aligned} t_c &= t/(n-1) \text{ or} \\ F_c &= F/(n-1)^2 \end{aligned} \tag{1}$$

(a general proof of this adjustment was provided by Ulrich & Miller, 2001).

Within the field of ERP research, the jackknife approach has been used for scoring onset latencies of the lateralized readiness potential (LRP; Miller et al., 1998; Ulrich & Miller, 2001), the P1, and the N1 (Jentzsch, Leuthold, & Ulrich, 2007). For LRP data this approach has been found to be significantly more accurate than scoring single-participant waveforms (Miller et al., 1998; Ulrich & Miller, 2001; but see Mordkoff & Gianaros, 2000) . Intuitively, one would expect an approach that works well for scoring LRP onset latency also to work well for scoring the latencies of other components of the ERP. In the absence of any clear theoretical understanding of the exact circumstances under which the jackknife approach works well, however, it is necessary to verify this intuition for other components on a case-by-case basis. In particular, there are important differences between N1 (auditory, visual), P3, N2pc, and frequency-related P3, on the one hand, and LRP, on the other hand, that might change the relative efficacy of the jackknife approach with these other components: First, the amplitudes of the N1 components are typically larger (at least with auditory stimuli) and the P3 component is clearly larger than LRP amplitudes. Second, at least for the N1 components, N2pc, and frequency-related P3, the onset and offset are more clearly defined. These two points make it somewhat easier to estimate latency differences for N1, N2pc, and P3 components than for LRPs. Because of these differences and others, it is desirable to compare single-participant and jackknife-based approaches while applying different scoring techniques for N1, N2pc, frequency-related P3, and P3 latency by means of simulations. Moreover, latency differences for all components could vary depending on the phenomenon that is studied, and they could be much smaller than the latency differences previously tested for the LRP (48 ms to 100 ms). Given that smaller latency differences (i.e., smaller effect sizes) would tend to make it more difficult to obtain significant differences, we chose, for each component, effect sizes that approached observed effect sizes in previously published empirical work for the respective component of interest.

Technique for Determining ERP Latency

Both the single-subject approach and the jackknife approach can be combined with several techniques for determining ERP latency. To determine whether the latency of an ERP component differs across two conditions, it is not necessary to know the true latency of the component in each condition; instead, it is sufficient to estimate the difference in latencies between the two conditions accurately (cf. Miller et al., 1998). To obtain a good estimate of this difference, it is necessary to choose an ERP onset latency criterion level large enough that it will rarely be crossed by chance in the average waveform being scored. ERP latencies are then measured as the first time point at which the ERP waves for both conditions fall below (for negative-going shifts like the N1 or the N2pc) or increase above (for positive-going shifts like the P3 or frequency-related P3) the criterion. The estimated latency difference, D, is the difference between these two time points.

In the remainder of this article, we report simulations evaluating various scoring techniques for estimating latency differences, using each scoring technique both with data from single participants and with the jackknife approach. To test a wide range of possible techniques, we included both absolute and relative criteria for detecting ERP onset. With the absolute criterion technique, ERP onset is defined as the time point at which the ERP amplitude reaches a constant prespecified value (e.g., − 0.5 μV for N1 or +1.0 μV for P3). With the relative criterion technique, ERP onset is defined as the time point at which the amplitude reaches a constant prespecified percentage of the peak value (e.g., 10% of the peak amplitude). Note that in these terms the peak latency is simply the point at which the 100% relative criterion is reached. We also included a technique defining latency in terms of a deviation from the value of the waveform during the baseline period (Osman, Bashore, Coles, Donchin, & Meyer, 1992). For this baseline deviation technique, the standard deviation during the baseline period is calculated to estimate the noise in the measured ERPs. Onset of the component is then determined as the first time point at which the ERP consistently falls below or increases above a criterion value set to a prespecified multiple (e.g., 2.0 times) of the standard deviation of baseline noise. Finally, we included the fractional area technique (Hansen & Hillyard, 1980; Luck, 2005). This technique defines the latency of the component as the first time point at which a certain percentage of the total area of the component has been reached. As shown in Figure 3, for example, the total area of an N1 component might be identified as the area of the ERP below − 2.0 μV in the time region from 0 to 150 ms after stimulus onset. The 50% fractional area latency measure would then be defined as the time point before which 50% of the total N1 area was observed. Note that two parameters must be chosen to use the fractional area latency measure: First, the percentage of the area that has to be reached and, second, the boundary from which the area is integrated. Setting the boundary differently from 0 μV makes sense both because noise can cause fluctuations around the 0-μV baseline and because many components do not start from a 0-μV level or do not finish at this level because of superposition with other components.

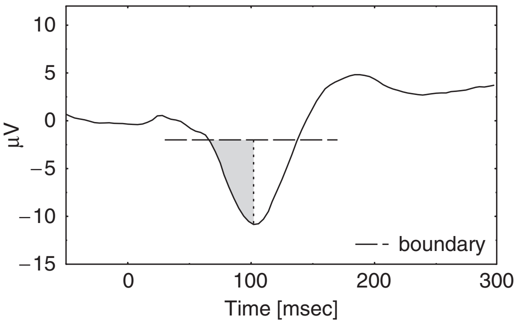

Figure 3. Visualization of the fractional area technique. Area of the component is defined as the area below (for positive components above) a chosen boundary (here − 2.0 μV) within a given time window. Latency is defined as the time point before which a prespecified percentage (here 50%, see gray-shaded area) of the total area was observed.

Each of these techniques can be used with different parameters. For example, the absolute criterion technique may determine latency at the time point at which the ERP amplitude reaches − 0.5 μV, − 1.5 μV, or − 5.0 μV. Likewise, the fractional area technique may determine latency at the time point at which 30% or 50% of the area under the curve is reached, relative to a boundary of − 0.5 μV or − 2.0 μV. To guide the choice of the parameters for the absolute criterion technique, the relative criterion technique, the baseline criterion technique, and the fractional area technique in future application of these techniques, we tested several parameter values for each technique (see description of the simulation protocol).

To evaluate the different procedures, we need to know the sampling distributions of the latency values produced by each procedure. Because the measurement procedures are complicated and the distributional properties of the underlying sources of noise are unknown, it is not possible to calculate these sampling distributions analytically. Therefore, we ran computer simulations (Monte Carlo simulations) to estimate these sampling distributions.

Of course, these estimated sampling distributions depend on the data sets that are used for the simulations. Unfortunately, it is not possible to predict whether a procedure that is preferable for one data set is suitable for other data sets as well. Therefore, we ran the computer simulations for five different ERP components that differ widely from each other on various dimensions.

In the following, the simulations for the N1 components (visual and auditory), the P3, the N2pc, and the frequency-related P3 are reported in separate parts. Thus, readers interested in the simulation results for a particular component may skip the alternative parts.

Simulations for the N1 Component

The N1 component is an often used measure of attention-related sensory processing (e.g., Eimer, 1995; Vogel & Luck, 2000). In particular, there has been considerable interest in effects of selective attention on the N1, and several studies have shown that the amplitude of the visual N1 response is larger for stimuli presented at an attended location relative to the response at unattended locations (Mangun, 1995). The auditory N1 response has also been shown to be modulated by selective attention (e.g., Hillyard, Hink, Schwent, & Picton, 1973; Woldorff et al., 1993). Reports regarding latency shifts are not as frequent, but, for example, the auditory N1 response has been shown to be delayed in repeated sound sequences (e.g., Dimitrijevic & Stapells, 2006; Sable, Low, Maclin, Fabiani, & Gratton, 2004).

General Simulation Protocol

To evaluate which procedure of analysis, that is, combination of approach (i.e., single-participant vs. jackknife-based), scoring technique (i.e., latency of peak, absolute criterion, relative criterion, baseline criterion, or fractional area), and parameter value yields the most accurate estimates of N1 latency differences, we ran a series of simulations with varying experimental conditions (e.g., sample size). Each simulation iterated a two-step process 1000 times, with each iteration representing one whole experiment (the simulation protocol resembles the protocol used by Miller et al., 1998). The two-step process of each iteration included the following: (a) A random set of data was generated to represent the outcome of a single experiment. (b) The generated data were analyzed with every combination of analysis approach and scoring technique.

To evaluate the accuracy and reliability of the different procedures for scoring N1 latency considered here, it is necessary to know the true size of the effect (i.e., the latency difference) underlying each simulated data set. Furthermore, the simulated data sets should be as realistic as possible (e.g., with respect to EEG noise) in order to get valid simulation results. To fulfill both of these requirements, we derived the simulated data sets from actual observed data sets using a method analogous to that employed by Miller et al. (1998). Specifically, we started with real EEG data from experiments in which the N1 component was similar in all experimental conditions. In each simulation, *n* participants were chosen randomly from the available pool of real participants. Then, a fixed number of trials were chosen from each participant's data set, and these trials were randomly subdivided into experimental and control trials. We let *t* indicate the number of trials in each condition. The entire EEG waveforms for all of the trials were shifted in time for the trials assigned to the experimental condition only; therefore, the size of this shift reflects the true effect size. The possible effect sizes that could be realized in the simulation were somewhat constrained by the 256-Hz sampling rate in the original data sets (i.e., the possible shifts were 3.90625 ms, 7.8125 ms, etc.), and we chose to shift the EEG data of the experimental trials 7.8125 ms, because this resembles the typical effect sizes for N1 latency differences (e.g., Jemel et al., 2003). Shifting the data points in time created a gap of two data points at the beginning of the baseline period in the EEG recordings. This two-sample gap was filled in with the original values for these two points, and these values were taken in reverse order to avoid creating a discontinuity.

To check whether the simulation results would be similar for different data sets, we used data from two rather different experiments, one measuring auditory evoked N1 at Cz, and one measuring visual evoked N1 at PO7 and PO8. The grand averages for both data sets are shown in Figure 4.

Data Set A (Figure 4, upper panel) involved measurement of auditory evoked N1 at Cz produced by a tone stimulus in a PRP experiment (for a detailed description, see Brisson & Jolicœur, 2007a). The full data set contained the data of 22 participants with at least 793 artifact-free trials per participant (average 967 trials). The baseline period was 50 ms prior to tone onset, and a recording epoch lasting until 300 ms after tone onset was used. The sampling rate was 256 Hz, and the EEG data were low-pass

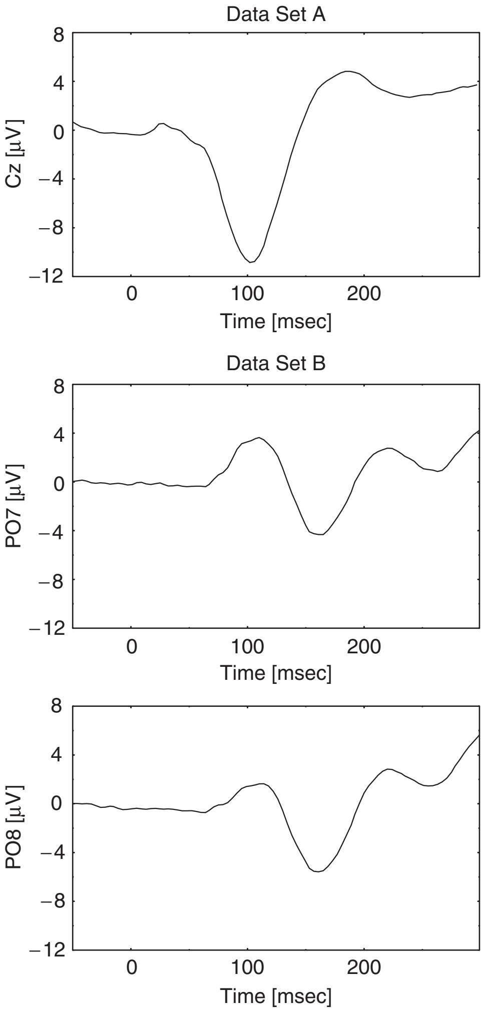

Figure 4. Grand average of Data Sets A and B that were used to generate the data for the simulations. Upper panel: auditory evoked N1 at Cz for Data Set A. Middle and lower panels: visually evoked N1 at PO7 and PO8 for Data Set B. (For better comparability of both data sets, only data within the range of – 50 ms to 300 ms from stimulus onset are depicted. However, for Data Set B, data from – 100 ms to 500 ms from stimulus onset were used).

filtered at 67 Hz and baseline corrected. Trials with eyeblinks (VEOG >80 μV), large horizontal eye movements (HEOG>35 μV), and within-trial deviations (i.e., difference between the maximum and minimum voltage values in an epoch) exceeding 80 μV at Cz were rejected.

Data Set B involved measurement of visual evoked N1 at PO7 (Figure 4, middle panel) and PO8 (Figure 4, lower panel) in response to squares that were presented in the lower visual field (for a detailed description, see De Beaumont, Brisson, Lassonde, & Jolicœur, 2007; only data of the nonconcussed athletic control group were considered). The set contained the data of 16 participants (2 of the original sample were not used because there were not enough artifact-free trials) with at least 289 artifact-free trials each (average 430). The baseline period consisted of the 100 ms before stimulus onset, and the recording epoch lasted until 500 ms after stimulus onset. The sampling rate was 256 Hz, and the data were low-pass filtered at 67 Hz and baseline corrected. Trials with eyeblinks (VEOG>80 μV), large horizontal eye movements (HEOG>35 μV), and within-trial deviations exceeding 80 μV at PO7 and/or PO8 were rejected.

When comparing the two data sets, it is obvious that N1 amplitude is larger in Data Set A than B, and in Data Set B, it is larger at PO8 than at PO7 (see Figure 4). Additionally, the visual evoked potentials in Data Set B reveal an initial positive deflection (i.e., P1) before the N1. Inspection of each participant's average waveform revealed that within Data Set A, the N1 amplitude is large for each participant. Within Data Set B, the N1 amplitude is generally smaller than in Data Set A, and at electrode PO7 some participants do not reveal any negative component in the N1 time window. Additionally, the N1 amplitudes are more variable in Data Set B than Data Set A. Thus, the signal-to-noise ratio is higher in Data Set A than B; moreover, within Data Set B, it is higher at electrode PO8 than PO7. Therefore, we expected that it would be easier to obtain statistically significant results for the artificially generated time shifts in the simulations with Data Set A than with Data Set B. Moreover, the large differences between data sets are an advantage for the present purposes, because they allow us to compare different scoring procedures under a wider range of conditions.

For each simulated experiment (i.e., iteration), the randomly selected experimental and control trials of the randomly chosen participants were analyzed with both the single-participant approach and the jackknife approach. The techniques for determining N1 onset latency were adjusted separately for the two data sets based on their somewhat different characteristics. For Data Set A, all techniques searched in the time window 0–150 ms after stimulus onset. The techniques to estimate latency were the peak latency; an absolute negative deflection falling below – 0.5, – 1.0, . . ., – 4.0 μV; a relative negative deflection falling below 10%, 30%, 50%, 70%, or 90% of the peak negative amplitude; and a baseline deviation of 2.0, 2.5, or 3.0 standard deviations. For the fractional area technique, parameters of 30%, 50%, or 70% of the area below the boundary were chosen, and these were combined with boundaries set to 0.0, – 1.0, or – 2.0 μV. For Data Set B, the techniques searched in the time window from 50 to 300 ms after stimulus onset. The techniques to estimate latency were the same as for Data Set A, except that the values for the absolute criterion technique were – 0.25, – 0.50, – 0.75, . . . , – 2.0 μV.

The parameters for the absolute criterion technique were chosen to be relatively small compared to the overall amplitude of the N1 to ensure that the criterion value was reached in most of the waveforms that were scored. If the criterion was not reached in the time window for scoring (e.g., for a single participant), however, the end of the time window (150 ms for Data Set A and 300 ms for Data Set B) was taken as the latency. Furthermore, if the criterion had already been reached at the beginning of the time window for scoring, the starting time of this window (i.e., 0 ms for Data Set A and 50 ms for Data Set B) was taken as the latency.

For all techniques other than the peak latency technique, the time point at which the criterion was reached was determined using linear interpolation. For example, suppose the absolute criterion was set to – 1.0 μV. Suppose further that a waveform had a value of – .98 μV at the time point of 62.50 ms and a value of – 1.02 μV at the next time point, 66.41 ms; in that case, the latency for reaching the absolute criterion would be calculated as 62.50+(– 1.0+0.98)/(– 1.02+0.98)* (66.41 – 62.50) = 64.46 ms.

Overview of the Simulations Results

Several different statistical indicators must be considered when judging which combination of approach (single participant or jackknife) and scoring technique provides the best measure of N1 latency differences. In the following, we first present results indicating how accurately the N1 latency differences (i.e., time shifts) were estimated (see Tables 1a and 1b). Then, we present simulations evaluating statistical power (i.e., the probability of obtaining significant latency differences, see Tables 2a, 2b, and 2c). Finally, we also report simulations conducted to estimate the Type I error rate of each procedure (i.e., the probability of obtaining a statistically significant difference when there was no true time shift in the simulated data, see Table 3).

Next, we report simulations for a between-subjects comparison. We present results regarding the estimation of latency differences and power to detect them in Tables 4a and 4b. The simulations to estimate Type I error rate for between-subjects comparisons are shown in Table 5.

The results are always listed in the same order in the tables: Results regarding the single-participant approach are shown to the left of results regarding the jackknife approach. The different techniques with the likewise applied parameters are listed line by line, where peak latency is shown first followed by absolute and relative criteria, baseline deviation criterion, and fractional area criterion. However, when describing the results, we do not follow this order. Instead we order the result sections according to the obtained outcome. First, we exclude as many procedures as possible by mentioning those techniques that are inferior irrespective of whether they are combined with the single-participants or the jackknife approach. Next, we state in detail which technique is preferable in general or just in combination with either the single-participant or the jackknife approach.

Table 1a. *Mean (M) and Standard Deviation (SD) in Milliseconds of the Estimated Differences (D) for Data Set A, Auditory N1 at Electrode Cz*[a]

	Single participants		Jackknife	
Criterion	*M*	*SD*	*M*	*SD*
Peak	7.79	1.80	7.98	2.36
Absolute criterion (μV)				
0.5	−0.08	9.20	3.66	27.04
1.0	2.22	10.82	7.30	22.18
1.5	4.36	10.91	7.90	8.17
2.0	5.93	9.92	7.95	4.25
2.5	7.09	8.13	7.72	2.67
3.0	7.42	6.86	7.37	2.11
3.5	7.46	5.69	7.16	1.96
4.0	7.59	4.73	7.12	1.88
Relative criterion (% maximum amplitude)				
10	−0.52	10.65	1.20	21.87
30	5.74	6.00	6.96	1.80
50	7.30	2.65	7.54	1.47
70	7.77	1.74	7.90	1.55
90	7.94	1.64	8.00	1.83
Baseline deviation (number of noise *SD*)				
2.0	−0.54	11.85	−1.23	29.67
2.5	0.25	11.06	1.22	26.70
3.0	1.41	10.18	3.38	22.46
Fractional area (% area/boundary)				
30/0	7.04	2.09	7.80	1.90
50/0	7.15	1.24	7.85	1.28
70/0	6.67	1.01	7.01	1.08
30/−1	7.51	1.72	8.03	1.27
50/−1	7.42	1.12	8.02	1.03
70/−1	6.92	1.00	7.17	1.02
30/−2	7.69	1.42	8.04	1.07
50/−2	7.54	1.08	8.04	0.96
70/−2	7.12	1.06	7.26	1.01

[a]The true mean difference was 7.8125 ms.

The main findings obtained in the simulations are summarized in the section "Discussion of N1 Results." Readers not interested in a detailed results description may continue there.

Estimation of 7.81-ms Effects

Tables 1a and 1b summarize the results of simulations that evaluate how accurately each procedure estimates N1 latency differences. The simulations in Table 1a were based on Data Set A, for the auditory N1 measured at the Cz electrode. The simulations in Table 1b were based on Data Set B, for the visual N1 observed at electrodes PO7 and PO8. In both simulations, $n = 12$ participants were chosen randomly without replacement from the available data pools (which contained 22 participants for Data Set A and 16 participants for Data Set B) for each simulated experiment. For each participant, 50 experimental and 50 control trials were randomly chosen (also without replacement) from all of the trials available for that participant. The experimental trials were shifted exactly 7.8125 ms. Each simulation included 1000 single experiments to estimate the differences (D) for the N1 latency in the experimental versus control conditions for each scoring procedure.

Means and standard deviations (SD) of the estimated differences (D) are listed in Tables 1a and 1b. Given that the true simulated difference was 7.8125 ms, a scoring procedure is better to the extent that it produces a mean value close to this value and also to the extent that it produces a smaller SD. Clearly, many of the combinations did very poorly. Both absolute and relative criterion techniques with small criterion parameter values tended to produce average difference estimates much smaller than the true value, as did the baseline deviation criterion with any number of noise SD. The absolute criterion technique performed somewhat better with larger criteria but still sometimes underestimated the true time shift (i.e., in Data Set A with the jackknife approach and in Data Set B with the single-participant approach).

Determining N1 latency with peak amplitude was quite accurate for Data Set A, but this technique yielded quite high SD for Data Set B, especially when combined with the single-participant approach. Using the relative criterion technique with a parameter of 50% or more provided accurate estimates with low standard deviations for both approaches for Data Set A, but for Data Set B the SD were high when using the single-participant approach. The single-participant approach combined with the fractional area technique tended to underestimate the true difference systematically—particularly for Data Set B. But, the combination of the fractional area technique with the jackknife approach had both accurate estimation and low SD when the area was determined with a negative boundary (i.e., -1.0 or -2.0). Interestingly, the jackknife approach was especially advantageous in Data Set B, that is, when the signal-to-noise ratio was lower.

To conclude, the peak latency technique, the relative criterion technique with parameters of 50% or more, and the fractional area technique with negative boundary were the most suitable for estimating N1 onset latency differences. The jackknife approach performed at least as well as the single-participant approach, and in fact its performance was definitely superior when the signal-to-noise ratio was lower (Data Set B, Table 1b).

Table 1b. *Mean (M) and Standard Deviation (SD) of Estimated Differences (D) for Data Set B, Visual N1 at Electrodes PO7 and PO8*[a]

	PO7				PO8			
	Single participants		Jackknife		Single participants		Jackknife	
Criterion	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Peak	6.69	11.36	7.96	3.96	7.34	6.83	7.96	3.52
Absolute criterion (μV)								
0.25	–1.24	21.75	–2.54	37.02	–.04	13.24	.46	19.70
0.5	–.90	21.00	–.11	45.04	.30	14.56	2.45	33.77
0.75	–.32	19.67	3.65	41.11	.64	15.35	4.54	44.78
1.0	.28	18.92	5.48	28.06	1.36	16.69	6.27	41.96
1.25	.50	18.57	6.92	13.45	2.36	17.82	6.00	26.71
1.5	.62	18.16	7.64	5.40	3.55	18.57	7.37	11.59
1.75	1.33	17.48	7.70	2.49	4.01	18.76	8.06	4.87
2.0	2.11	16.93	7.70	2.34	4.73	18.11	8.00	2.35
Relative criterion (% maximum amplitude)								
10	.29	22.73	–2.26	43.49	–.18	14.10	–1.43	34.77
30	2.87	23.09	7.22	11.29	2.77	13.80	7.52	7.82
50	5.18	21.68	7.49	1.51	5.32	11.64	7.76	1.72
70	6.17	21.02	7.05	1.50	6.90	9.49	7.24	1.75
90	6.50	20.79	7.60	2.66	7.20	8.51	7.53	2.06
Baseline deviation (number of noise *SD*)								
2.0	.54	18.73	–3.38	46.86	.23	17.48	–6.25	39.81
2.5	1.35	18.86	–1.66	44.80	1.00	19.00	–7.25	44.71
3.0	2.18	19.67	.99	40.47	1.79	20.25	–5.85	44.81
Fractional area (% area/boundary)								
30/0	4.24	9.89	6.90	1.88	5.23	7.90	6.58	2.21
50/0	4.73	10.26	7.57	1.57	6.27	6.31	7.60	1.59
70/0	5.32	11.01	7.68	1.60	6.88	5.67	7.72	1.35
30/–1	5.14	8.47	7.73	1.17	6.34	9.33	7.42	1.20
50/–1	5.28	8.28	7.92	1.24	6.78	8.35	7.95	1.15
70/–1	5.46	8.82	7.89	1.42	7.04	8.12	7.94	1.26
30/–2	5.43	7.25	7.81	1.46	6.81	10.04	7.62	1.20
50/–2	5.47	7.10	7.90	1.54	7.00	9.94	7.97	1.21
70/–2	5.42	8.07	7.88	1.72	6.98	10.21	7.96	1.35

[a]The true mean difference was 7.8125 ms.

Power to Detect 7.81-ms Effects

In addition to evaluating the means and *SD* of the different procedures to estimate N1 latency differences, it is also important to consider the statistical power obtained with each procedure (i.e., the probability of obtaining statistically significant evidence of a latency difference due to the simulated time shift). Therefore, we ran some further simulations in which we calculated separate *t* tests for each simulated experiment. Significance level (α – error) was set to $p = .05$ for each two-tailed *t* test. For Data Set A, the simulations were run with the number of participants, $n = 8$, 12, and 20, and with the number of trials, $t = 30$, 50, and 70. For Data Set B, simulations were run with $n = 8$ and 12 and with $t = 30$, 50, and 70. Each simulation again included 1000 experiments; the proportion of these yielding significant *t* tests is an estimate of the power to detect the 7.81-ms time shift.

The estimated power values are listed in Table 2a for Data Set A, in Table 2b for Data Set B and electrode PO7, and in Table 2c for Data Set B and electrode PO8.

As expected, power increased for all procedures with the number of participants, the number of trials, and the signal-to-noise ratio, which seems to be the highest in Data Set A, followed by the PO8 electrode of Data Set B, and seems to be the lowest for the PO7 electrode of Data Set B.

The three different data sets revealed similar results with some procedures. Power was rather low for both the single-participant and the jackknife approach when using the absolute criterion technique, the relative criterion technique with parameters below 50%, or any baseline deviation technique. In contrast, power was reasonably high in all cases using the fractional area technique with a negative boundary, and it was especially high when combining the fractional area technique with the jackknife approach (this was especially evident for Data Set B, in which the N1 had a low signal-to-noise ratio). Interestingly, with other procedures, the pattern of power differed depending on the data set. For Data Set A, power was large for both approaches when using the relative criterion technique with parameters 50%, 70%, or 90%. When using the peak latency technique, however, the single-participant approach did very well, but power was low for the jackknife approach. In contrast, for Data Set B, power was high when we used the jackknife approach combined with the relative criterion techniques with parameters in the range of 50% to 90%, and power was relatively low for the single participant approach and for the peak latency technique with either approach (especially for the PO7 electrode).

Overall, the jackknife approach combined with the fractional area technique was clearly the best choice, because it had high power for both data sets. In comparison, the more standard procedure of scoring peak latency for single participants was only suitable for data sets with high signal-to-noise ratio. If the data are noisy, the power to detect latency differences drops dramatically with the single-participant peak latency procedure. Interestingly, combining the jackknife approach with the peak latency technique did not increase power, but, in fact, compared to the single subject approach, even reduced it.

Table 2a. *Power to Detect a Mean Latency Difference of 7.8125 ms as a Function of Method and Criterion, Depending on Number of Trials and Number of Participants for Data Set A, Auditory N1 at Electrode Cz*

	8 participants						12 participants						20 participants					
	Single participants			Jackknife			Single participants			Jackknife			Single participants			Jackknife		
Criterion	30 trials	50 trials	70 trials	30 trials	50 trials	70 trials	30 trials	50 trials	70 trials	30 trials	50 trials	70 trials	30 trials	50 trials	70 trials	30 trials	50 trials	70 trials
Peak	.725	.851	.927	.237	.263	.265	.885	.966	.990	.282	.294	.287	.986	1.00	1.00	.374	.381	.402
Absolute criterion (μV)																		
0.5	.017	.011	.016	.039	.034	.039	.027	.022	.018	.043	.045	.034	.019	.022	.022	.072	.061	.062
1.0	.027	.030	.024	.056	.070	.083	.038	.046	.028	.080	.107	.117	.031	.042	.039	.085	.117	.155
1.5	.046	.055	.053	.092	.157	.230	.056	.059	.052	.162	.224	.283	.048	.082	.082	.256	.326	.402
2.0	.060	.091	.097	.180	.327	.412	.073	.101	.114	.339	.501	.602	.090	.122	.153	.571	.785	.904
2.5	.073	.132	.184	.302	.470	.624	.100	.150	.206	.539	.742	.840	.119	.202	.272	.812	.948	.990
3.0	.113	.184	.289	.380	.595	.750	.127	.258	.334	.653	.857	.923	.166	.311	.457	.895	.986	1.00
3.5	.164	.281	.434	.428	.667	.797	.174	.360	.504	.691	.887	.938	.223	.443	.634	.899	.988	1.00
4.0	.204	.384	.564	.480	.715	.838	.231	.503	.659	.724	.903	.954	.317	.564	.804	.915	.989	1.00
Relative criterion (% maximum amplitude)																		
10	.024	.021	.023	.040	.049	.077	.030	.024	.011	.064	.069	.093	.023	.027	.016	.067	.081	.110
30	.131	.223	.349	.498	.692	.821	.129	.284	.404	.741	.916	.955	.143	.281	.472	.945	.996	1.00
50	.495	.731	.830	.727	.894	.954	.568	.804	.912	.915	.978	1.00	.655	.892	.968	.994	1.00	1.00
70	.728	.890	.936	.747	.898	.950	.855	.966	.994	.929	.989	1.00	.940	.997	1.00	.991	1.00	1.00
90	.796	.883	.948	.589	.752	.850	.932	.985	.997	.779	.923	.970	.988	.999	1.00	.942	.995	1.00
Baseline deviation (number of noise *SD*)																		
2.0	.027	.024	.024	.040	.042	.047	.029	.026	.018	.055	.052	.057	.017	.018	.022	.071	.072	.067
2.5	.029	.028	.029	.034	.047	.043	.033	.033	.026	.059	.066	.059	.028	.023	.028	.078	.070	.093
3.0	.043	.044	.048	.041	.057	.061	.040	.038	.033	.054	.069	.070	.039	.040	.038	.083	.099	.113
Fractional area (% area/boundary)																		
30/0	.511	.676	.806	.580	.770	.862	.674	.855	.921	.788	.929	.977	.827	.957	.992	.941	.993	1.00
50/0	.837	.949	.990	.873	.962	.995	.961	.998	.999	.986	.999	1.00	.997	1.00	1.00	1.00	1.00	1.00
70/0	.923	.989	.999	.912	.978	.997	.994	.999	1.00	.983	.999	1.00	1.00	1.00	1.00	1.00	1.00	1.00
30/ – 1	.662	.853	.915	.851	.954	.990	.817	.945	.985	.979	.997	1.00	.935	.996	.999	1.00	1.00	1.00
50/ – 1	.892	.978	.995	.951	.994	1.00	.979	.999	1.00	1.00	1.00	1.00	.998	1.00	1.00	1.00	1.00	1.00
70/ – 1	.939	.991	.999	.942	.990	.999	.994	.999	1.00	.992	.999	1.00	1.000	1.00	1.00	1.00	1.00	1.00
30/ – 2	.781	.928	.977	.942	.989	.999	.908	.984	.997	.997	1.00	1.00	.979	.999	1.00	1.00	1.00	1.00
50/ – 2	.932	.989	.999	.984	.998	1.00	.990	.998	1.00	.999	1.00	1.00	.999	1.00	1.00	1.00	1.00	1.00
70/ – 2	.941	.989	1.00	.954	.992	.999	.994	.998	1.00	.995	1.00	1.00	.999	1.00	1.00	1.00	1.00	1.00

Table 2b. *Power to Detect a Mean Latency Difference of 7.8125 ms as a Function of Method and Criterion, Depending on Number of Trials and Number of Participants for Data Set B, Visual N1 at Electrode PO7*

	8 participants						12 participants					
	Single participants			Jackknife			Single participants			Jackknife		
Criterion	30 trials	50 trials	70 trials	30 trials	50 trials	70 trials	30 trials	50 trials	70 trials	30 trials	50 trials	70 trials
Peak	.169	.240	.320	.103	.118	.136	.159	.194	.246	.134	.134	.118
Absolute criterion (μV)												
0.25	.007	.005	.002	.031	.021	.036	.006	.004	.009	.029	.020	.027
0.5	.007	.005	.007	.053	.058	.081	.007	.007	.008	.058	.077	.106
0.75	.010	.007	.014	.084	.124	.198	.007	.009	.009	.145	.250	.304
1.0	.010	.007	.019	.133	.235	.342	.016	.008	.024	.273	.518	.622
1.25	.013	.018	.025	.204	.336	.509	.018	.022	.022	.409	.716	.840
1.5	.013	.014	.036	.259	.447	.624	.024	.023	.022	.511	.823	.900
1.75	.015	.022	.043	.312	.504	.665	.020	.024	.036	.561	.819	.902
2.0	.017	.024	.055	.311	.495	.659	.021	.026	.045	.575	.786	.887
Relative criterion (% maximum amplitude)												
10	.010	.009	.007	.055	.061	.081	.008	.012	.011	.052	.058	.070
30	.027	.031	.059	.364	.540	.700	.028	.035	.021	.582	.845	.928
50	.075	.099	.146	.578	.752	.880	.058	.058	.052	.850	.961	.985
70	.090	.133	.149	.536	.718	.811	.055	.077	.067	.808	.923	.975
90	.095	.130	.152	.271	.369	.495	.066	.090	.078	.365	.498	.610
Baseline deviation (number of noise *SD*)												
2.0	.018	.010	.015	.078	.085	.101	.023	.019	.020	.083	.081	.071
2.5	.017	.020	.028	.090	.115	.142	.021	.015	.026	.111	.124	.102
3.0	.021	.021	.041	.120	.140	.183	.027	.028	.029	.144	.191	.179
Fractional area (% area/boundary)												
30/0	.127	.194	.257	.447	.573	.694	.129	.166	.204	.621	.790	.883
50/0	.120	.177	.193	.574	.716	.814	.128	.142	.174	.824	.942	.974
70/0	.109	.176	.226	.481	.648	.742	.106	.148	.217	.779	.895	.950
30/ – 1	.200	.329	.414	.733	.861	.941	.173	.289	.335	.960	.993	1.00
50/ – 1	.181	.273	.337	.766	.883	.944	.154	.277	.304	.974	.998	1.00
70/ – 1	.153	.272	.346	.633	.804	.882	.141	.243	.359	.926	.995	.998
30/ – 2	.299	.458	.547	.686	.788	.858	.243	.413	.459	.908	.970	.987
50/ – 2	.226	.365	.451	.690	.799	.866	.204	.363	.421	.912	.979	.994
70/ – 2	.182	.276	.341	.613	.733	.834	.158	.271	.290	.867	.963	.988

Table 2c. *Power to Detect a Mean Latency Difference of 7.8125 ms as a Function of Method and Criterion, Depending on Number of Trials and Number of Participants for Data Set B, Visual N1 at Electrode PO8*

	8 participants						12 participants					
	Single participants			Jackknife			Single participants			Jackknife		
Criterion	30 trials	50 trials	70 trials	30 trials	50 trials	70 trials	30 trials	50 trials	70 trials	30 trials	50 trials	70 trials
Peak	.310	.509	.593	.117	.135	.113	.337	.464	.603	.126	.125	.124
Absolute criterion (μV)												
0.25	.003	.003	.007	.015	.020	.009	.013	.011	.008	.019	.010	.002
0.5	.005	.007	.009	.026	.029	.033	.015	.011	.014	.027	.034	.026
0.75	.014	.014	.021	.040	.057	.053	.025	.018	.016	.071	.078	.094
1.0	.020	.020	.028	.068	.143	.187	.024	.026	.024	.164	.224	.318
1.25	.013	.023	.026	.147	.259	.375	.023	.037	.034	.305	.496	.690
1.5	.020	.032	.033	.228	.396	.556	.029	.039	.041	.484	.745	.883
1.75	.026	.032	.038	.303	.512	.690	.027	.042	.041	.586	.847	.928
2.0	.023	.033	.063	.374	.570	.740	.032	.041	.051	.635	.847	.929
Relative criterion (% maximum amplitude)												
10	.016	.011	.015	.021	.033	.034	.016	.021	.015	.023	.038	.023
30	.022	.045	.069	.397	.604	.722	.040	.053	.067	.681	.882	.953
50	.095	.181	.322	.559	.735	.859	.092	.174	.256	.760	.906	.960
70	.207	.386	.523	.501	.659	.786	.148	.309	.391	.669	.836	.926
90	.280	.427	.548	.364	.463	.575	.232	.350	.483	.547	.701	.811
Baseline deviation (number of noise *SD*)												
2.0	.023	.016	.028	.045	.043	.032	.031	.015	.022	.049	.033	.015
2.5	.026	.026	.044	.054	.065	.055	.034	.031	.041	.091	.064	.041
3.0	.027	.027	.040	.069	.105	.088	.041	.034	.029	.133	.111	.074
Fractional area (% area/boundary)												
30/0	.141	.278	.382	.375	.517	.628	.150	.279	.401	.518	.708	.830
50/0	.307	.465	.535	.649	.816	.886	.304	.466	.495	.880	.969	.987
70/0	.292	.488	.587	.721	.871	.924	.334	.562	.672	.937	.993	.999
30/ – 1	.182	.314	.409	.701	.865	.929	.167	.261	.291	.907	.983	.995
50/ – 1	.316	.438	.531	.834	.950	.977	.289	.368	.418	.985	1.00	1.00
70/ – 1	.281	.418	.516	.830	.937	.979	.282	.394	.459	.969	.999	1.00
30/ – 2	.236	.404	.499	.806	.912	.962	.181	.308	.354	.973	.995	.999
50/ – 2	.291	.410	.500	.844	.937	.980	.215	.317	.364	.983	.999	.999
70/ – 2	.247	.366	.460	.798	.908	.963	.198	.306	.344	.949	.996	1.00

Table 3. *Type I Error as a Function of Method and Criterion, Depending on Significance Level (.05 vs. .01) for 12 Participants for Data Set A, Auditory N1 at Cz Electrode*

	Single participants						Jackknife					
	30 trials		50 trials		70 trials		30 trials		50 trials		70 trials	
Criterion	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01
Peak	.017	.004	.026	.002	.014	.004	.037	.037	.031	.031	.024	.024
Absolute criterion (μV)												
0.5	.023	.001	.022	.003	.027	.002	.063	.057	.048	.043	.029	.026
1.0	.029	.003	.035	.009	.026	.006	.058	.045	.036	.025	.025	.018
1.5	.020	.007	.028	.004	.029	.005	.035	.022	.020	.009	.016	.008
2.0	.033	.009	.022	.003	.027	.005	.021	.004	.010	.002	.013	.006
2.5	.032	.007	.020	.001	.021	.001	.012	.005	.009	.001	.014	.002
3.0	.022	.001	.016	.000	.014	.000	.015	.004	.015	.003	.015	.000
3.5	.020	.001	.013	.001	.008	.000	.017	.002	.015	.005	.019	.001
4.0	.013	.002	.010	.001	.007	.001	.017	.002	.021	.008	.017	.002
Relative criterion (% maximum amplitude)												
10	.025	.003	.036	.006	.022	.002	.053	.037	.031	.020	.018	.011
30	.019	.003	.007	.001	.011	.002	.018	.004	.013	.001	.011	.003
50	.013	.002	.012	.001	.016	.002	.023	.003	.015	.003	.016	.002
70	.018	.001	.012	.002	.016	.000	.018	.004	.018	.005	.021	.003
90	.026	.003	.020	.003	.023	.004	.022	.007	.014	.003	.012	.001
Baseline deviation (number of noise *SD*)												
2.0	.028	.006	.030	.002	.027	.009	.051	.038	.048	.042	.043	.038
2.5	.025	.005	.027	.004	.031	.007	.039	.026	.033	.023	.034	.022
3.0	.027	.003	.026	.002	.024	.000	.031	.018	.023	.015	.032	.017
Fractional area (% area/boundary)												
30/0	.020	.003	.022	.001	.029	.003	.027	.003	.015	.003	.017	.002
50/0	.026	.003	.023	.001	.023	.003	.021	.005	.014	.002	.017	.001
70/0	.022	.001	.024	.005	.021	.002	.022	.004	.014	.003	.017	.004
30/ – 1	.024	.001	.019	.001	.022	.003	.023	.004	.015	.003	.019	.003
50/ – 1	.026	.004	.017	.003	.024	.001	.019	.004	.017	.003	.017	.002
70/ – 1	.018	.001	.021	.004	.027	.003	.021	.003	.015	.002	.019	.001
30/ – 2	.025	.001	.015	.000	.021	.003	.025	.003	.019	.002	.012	.001
50/ – 2	.024	.003	.017	.004	.024	.004	.017	.005	.016	.002	.027	.000
70/ – 2	.016	.004	.023	.003	.023	.004	.018	.004	.013	.002	.018	.001

Type I Error

We also ran simulations to evaluate the Type I error rates of the different procedures of analysis. Clearly, a desirable method should produce not only accurate measurements of the true latency difference (i.e., yield the correct mean and low *SD*) and high statistical power in testing for significant differences, but it should also produce an appropriately low Type I error rate. The simulation protocol was identical to that used in the simulations examining power, except that experimental trials were not shifted along the time axis (i.e., the null hypothesis of no onset latency difference was in fact true). As in the power simulations, two-tailed t tests were calculated for each single experiment with significance levels of $p = .05$ and $p = .01$, and the numbers of statistically significant results were tabulated. These simulations were run using Data Set A and Data Set B with $n = 12$ participants and $t = 30$, 50, and 70 trials.

Table 3 lists the estimated Type I error proportion depending on significance level and number of trials for Data Set A. Results for Data Set B were similar. For reasons of brevity, we do not present tables listing Type I error for Data Set B. All of the procedures produced Type I error rates less than or equal to the nominal values of $p = .05$ and $p = .01$. Thus, inflation of Type I error rate is not a concern with any of the analysis procedures.

Between-Subject Comparisons

ERP latency differences are often investigated in within-subjects designs. However, there are also cases when it is interesting to compare N1 latencies across different groups. For example, one might be interested in comparing N1 latencies across different age groups (e.g., Curran, Hills, Patterson, & Strauss, 2001), patient populations versus a control group (e.g., Tachibana, Aragane, Miyata, & Sugita, 1997), left- versus right-handers (e.g., Alexander & Polich, 1995), and so on. Therefore, we also ran simulations to compare the jackknife and single-participant approaches with various onset scoring techniques for between-subjects designs.

For this purpose, the simulation protocol was changed to implement a between-subjects comparison. In each simulation step, 16 participants were chosen randomly from Data Set A. Half of the randomly chosen participants were assigned to the control group, and the other half were assigned to the experimental group. For Data Set B data, the 16 participants were randomly divided into a control and an experimental group. From the data pool of each participant, 50 trials were chosen randomly. Then the entire EEG waveforms for all of the trials of the experimental group were shifted 7.8125 ms in time. After this shift, the data for the experimental and control groups were analyzed with the single-participant and jackknife approaches using each of the latency estimation techniques, as in the previous simulations. The obtained latency values were compared with between-subjects t tests with a two-tailed significance level of $p = .05$. The t values for the jackknife approach were adjusted according to Equation 1 with $n =$ number of participants per group (i.e., $n = 8$).

The main simulation results are shown in Tables 4a and 4b. Again the absolute criterion technique, the relative criterion technique with parameters below 50%, or any baseline deviation technique were clearly inferior irrespective of whether they were combined with the single-participant or the jackknife approach. For Data Set A (Table 4a), accurate estimates of the mean differences combined with satisfactorily low standard deviations were obtained with the single-participant approach with the peak latency technique. The relative criterion technique with parameters of 50% and higher and the fractional area technique for both approaches also provided accurate estimates of the mean differences and low standard deviations. Likewise, the power to detect latency differences was highest for these procedures. Although there was a slight advantage for the jackknife approach when the simulations were based on Data Set A, the gains were modest. In contrast, for Data Set B (Table 4b), estimates of the mean differences were most accurate and power to detect latency differences was highest for the jackknife approach combined either with the relative criterion technique with parameters of 50% and higher or with the fractional area technique.

However, power to detect latency differences in between-subjects designs is generally much lower than in within-subject designs. To increase power, one should therefore increase the sample sizes in the experimental and control conditions if possible. And of course, one should be very cautious in interpreting null results if a between-subjects comparison does not reveal significant group differences.

To obtain a complete picture of between-subjects comparisons we also ran the simulations based on Data Sets A and B without a time shift to obtain an estimate of Type I error. The results were again satisfactory: Type I error was not increased in any of the procedures (see Table 5).

Discussion of Visual and Auditory N1 Results

The simulations for within-subject as well as between-subjects comparisons show that the standard procedure for evaluating N1 latency effects—namely, comparing the peak latencies for single participants—is not generally the most efficient one. If the signal-to-noise ratio is high (as in Data Set A), this procedure estimates and detects latency differences quite well. If the signal-to-noise ratio is lower (as in Data Set B), however, then the power to detect latency differences using single-participant approaches is low, and the jackknife approach combined with the fractional area technique is significantly better.

The jackknife approach estimates and detects latency differences satisfactorily for both data sets when we use the relative criterion technique with parameters of 50% and higher and when we use the fractional area technique. Indeed, when considering both data sets, the most accurate estimates of mean differences and the highest power values were observed for the jackknife approach with the fractional area technique with the parameter of 50% of the area and a boundary of -1.0 μV. With these parameters, no other technique outperformed the jackknife approach combined with the fractional area technique, including any version of the peak latency technique. On the basis of these findings, therefore, our overall recommendation is to use the jackknife approach with the fractional area technique with the parameters 50% area and a boundary of -1.0 μV to compare N1 latency differences in both within-subject and between-subjects designs.

Simulations for the P3 Component

In this portion of the article the same methods that were tested for the N1 components were applied to the P3 component. As mentioned in the Introduction, it is important to estimate the various latency estimation procedures separately with this component (and not just to generalize the results based on the N1 simulations) because of numerous differences between these two components.

The P3 is a large positive ERP component that generally peaks at around 300–400 ms following presentation of a task-relevant stimulus in any modality. It has a broad scalp distribution, generally maximal at central-parietal electrode sites (i.e., Pz when measured with a mastoid reference). The P3 component is sensitive to the probability of occurrence of the target in an attended stream of nontargets (Duncan-Johnson & Donchin, 1977). However, it does not seem to be the probability of the physical stimulus per se that matters, but rather the probability of the task-defined stimulus category (see Donchin & Coles, 1988; Kutas, McCarthy, & Donchin, 1977, Vogel, Luck, & Shapiro, 1998). Although there is still some disagreement (e.g., Verleger, 1988), the processes underlying the P3 are often thought to be associated with "context updating" (Donchin, 1981) or encoding into short-term memory. Because the P3 amplitude is sensitive to the probability of the class-defined stimulus category, it logically follows that some part of the P3 cannot be generated before the stimulus has been categorized. Moreover, the P3 latency (as well as its amplitude) is relatively insensitive to factors that influence response-selection processes, such as stimulus–response compatibility (Magliero, Bashore, Coles, & Donchin, 1984). On the basis of such observations, it has been hypothesized that P3 latency can be taken as an electrophysiological measure of the duration of stimulus-evaluation processes, in contrast to response processes (Coles, Smid, Scheffers, & Otten, 1995; Duncan-Johnson, 1981; Kutas et al., 1977; Leuthold & Sommer, 1998; McCarthy & Donchin, 1981; for a review and critique, see Verleger, 1997).

The P3 component differs from the N1 regarding several respects: Both the signal and the noise are larger for the P3 than for the N1 components examined in the foregoing sections. Thus, it is difficult to foresee whether estimates of P3 latency differences will be more or less accurate. Typical effects of experimental manipulations on latency are larger for P3 than N1 components, which should make it easier to obtain latency differences. Furthermore, the P3 component is broader (i.e., more extended in time). Therefore it is especially questionable whether the peak latency technique will still do a good job.

General Simulation Protocol

The general protocol for simulations examining P3 latency differences was similar to the protocol for simulations of N1 latency differences. Because experimental effects on latency tend to be somewhat larger for P3 than for N1 components, we decided to shift the EEG data of the experimental trials by 15.625 ms. This reflects a rather small P3 latency effect size but still lies in the range of observed effect sizes (e.g., Callaway, Halliday, Naylor, & Schechter, 1985; Mulder, 1984; Verleger, Neukater, Kompf, & Vieregge, 1991).

For the simulations, we took data from an experiment with visual stimulation in which the P3 was recorded at the Pz electrode (for a detailed description of the experiment, see De Beaumont et al., 2007; only data of the frequent condition in the

Table 4a. *Mean (M) and Standard Deviation (SD) of Estimated Differences (D) and the Power (1 – β) to Detect Latency Differences of 7.8125 ms in a Between-Subjects Design for Data Set A, Auditory N1 at Electrode Cz*

	Single participants			Jackknife		
Criterion	*M*	*SD*	1 – *β*	*M*	*SD*	1 – *β*
Peak	7.52	3.94	.441	7.76	4.28	.225
Absolute criterion (μV)						
0.5	– .66	13.01	.019	.68	34.47	.049
1.0	1.11	14.83	.028	6.02	29.65	.079
1.5	3.91	15.17	.036	7.34	15.24	.123
2.0	6.17	14.48	.065	7.59	8.17	.178
2.5	7.16	12.33	.071	7.53	5.57	.253
3.0	7.54	10.47	.115	7.32	4.55	.283
3.5	7.55	8.79	.156	7.18	4.24	.296
4.0	7.54	7.33	.186	7.17	4.17	.309
Relative criterion (% maximum amplitude)						
10	– 1.77	14.84	.019	– 1.52	29.97	.056
30	5.31	9.69	.126	6.70	4.01	.308
50	7.02	4.55	.392	7.27	3.65	.467
70	7.50	3.94	.461	7.65	4.01	.422
90	7.62	3.99	.449	7.60	4.39	.351
Baseline deviation (number of noise *SD*)						
2.0	– 2.17	15.47	.019	– 2.11	32.36	.042
2.5	– .72	14.36	.019	– .16	28.28	.039
3.0	.71	12.97	.030	1.99	23.99	.051
Fractional area (% area/boundary)						
30/0	6.80	4.25	.375	7.44	4.05	.401
50/0	6.89	3.58	.491	7.54	3.68	.493
70/0	6.43	3.72	.390	6.83	3.69	.386
30/ – 1	7.29	3.85	.477	7.79	3.60	.540
50/ – 1	7.16	3.58	.499	7.79	3.66	.514
70/ – 1	6.68	3.82	.383	7.05	3.80	.388
30/ – 2	7.47	3.61	.513	7.85	3.48	.564
50/ – 2	7.30	3.59	.494	7.85	3.66	.516
70/ – 2	6.87	3.90	.386	7.16	3.81	.392

Table 4b. *Mean (M) and Standard Deviation (SD) of Estimated Differences (D) and the Power (1 – β) to Detect Latency Differences of 7.8125 ms in a Between-Subjectd Design for Data Set B, Visual N1 at Electrodes PO7 and PO8*

	PO7						PO8					
	Single participants			Jackknife			Single participants			Jackknife		
Criterion	*M*	*SD*	1 – *β*	*M*	*SD*	1 – *β*	M	*SD*	1 – *β*	M	*SD*	1 – *β*
Peak	7.51	19.73	.056	8.28	7.76	.155	8.21	10.71	.097	7.92	7.09	.150
Absolute criterion (μV)												
0.25	– 3.41	39.45	.011	– 4.28	50.65	.052	– 1.37	20.47	.009	– .78	24.62	.016
0.5	– 3.18	41.91	.014	– 3.73	59.35	.076	– .88	23.22	.012	– 1.48	39.85	.041
0.75	– 3.15	43.21	.015	– 2.25	56.86	.078	– .47	25.57	.020	– 1.69	52.25	.072
1.0	– 2.38	44.13	.020	.61	45.41	.067	– .36	27.32	.021	.82	52.49	.080
1.25	– 2.04	45.01	.019	3.54	32.24	.080	.35	29.54	.026	2.71	42.22	.059
1.5	– 1.92	45.00	.027	5.50	23.39	.078	.14	31.61	.026	5.02	29.34	.066
1.75	– 2.28	44.23	.026	6.35	20.49	.078	.91	32.94	.028	6.95	18.47	.085
2.0	– 1.55	43.44	.028	6.47	23.80	.060	.97	34.11	.030	7.46	12.82	.095
Relative criterion (% maximum amplitude)												
10	– 1.52	36.31	.012	– 3.45	56.96	.066	– 1.40	20.82	.018	– 3.00	41.17	.048
30	2.40	33.65	.030	6.78	25.16	.126	2.92	22.23	.026	6.68	21.40	.099
50	5.13	30.76	.047	7.54	8.06	.135	6.08	18.98	.060	7.63	7.59	.156
70	6.11	28.41	.044	7.26	6.26	.139	7.64	15.98	.069	7.19	6.66	.123
90	6.97	26.67	.043	7.60	6.83	.189	7.77	13.10	.076	7.25	6.09	.182
Baseline deviation (number of noise *SD*)												
2.0	– 1.68	43.03	.026	– 2.89	57.21	.063	.40	31.80	.025	– 5.11	49.35	.058
2.5	– .97	42.78	.026	– 1.70	52.44	.054	1.50	33.02	.028	– 4.68	51.90	.061
3.0	.32	40.98	.024	– .12	45.87	.052	2.64	34.74	.031	– 3.00	48.44	.051
Fractional area (% area/boundary)												
30/0	6.07	20.58	.043	6.97	5.60	.181	6.30	14.07	.058	6.63	6.62	.116
50/0	6.86	22.20	.049	7.79	5.50	.176	7.42	11.19	.093	7.42	5.69	.194
70/0	7.76	25.87	.047	8.09	7.49	.126	7.83	10.65	.120	7.60	5.87	.194
30/ – 1	6.80	22.57	.048	7.76	5.57	.249	7.48	14.72	.057	7.22	5.83	.155
50/ – 1	7.53	24.73	.055	8.21	6.00	.223	8.23	13.72	.076	7.74	5.67	.200
70/ – 1	8.24	28.67	.044	8.37	7.13	.188	8.36	13.92	.097	7.78	5.90	.195
30/ – 2	7.53	23.28	.044	7.94	13.03	.198	7.18	17.28	.046	7.34	5.93	.177
50/ – 2	8.04	25.39	.053	8.36	14.51	.209	7.74	17.75	.055	7.78	5.76	.202
70/ – 2	8.81	29.64	.040	8.58	16.32	.167	7.93	18.70	.064	7.83	5.94	.191

Table 5. *Type I Error as a Function of Method and Criterion, Depending on Significance Level (.05 vs. .01) for a Between-Subjects Comparison Based on 16 Participants with 50 trials for Data Set A, Auditory N1 at Electrode Cz, and Data Set B, Visual N1, at Electrodes PO7 and PO8*

	Single participants						Jackknife					
	Set A: Cz		Set B: PO7		Set B: PO8		Set A: Cz		Set B: PO7		Set B: PO8	
Criterion	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01
Peak	.024	.003	.013	.000	.026	.004	.035	.033	.048	.028	.039	.026
Absolute criterion (μV)												
0.25	.027	.006	.012	.000	.009	.000	.066	.062	.062	.061	.007	.005
0.5	.028	.008	.015	.002	.009	.001	.077	.058	.094	.094	.032	.030
0.75	.028	.007	.016	.003	.013	.002	.055	.028	.088	.087	.078	.075
1.0	.032	.006	.022	.002	.012	.004	.027	.009	.042	.041	.081	.080
1.25	.024	.004	.027	.005	.013	.004	.012	.000	.030	.022	.048	.044
1.5	.028	.003	.029	.007	.011	.002	.018	.000	.024	.013	.030	.020
1.75	.019	.001	.029	.008	.014	.001	.026	.003	.026	.012	.021	.009
2.0	.016	.000	.025	.006	.015	.002	.022	.004	.022	.008	.026	.009
Relative criterion (% maximum amplitude)												
10	.027	.006	.015	.002	.015	.001	.065	.052	.107	.107	.047	.047
30	.014	.000	.021	.002	.016	.002	.016	.003	.035	.016	.027	.009
50	.017	.003	.023	.000	.028	.006	.017	.004	.029	.010	.040	.017
70	.026	.007	.020	.000	.020	.002	.027	.006	.021	.005	.034	.007
90	.025	.007	.015	.000	.021	.003	.027	.007	.062	.029	.030	.012
Baseline deviation (number of noise *SD*)												
2.0	.018	.004	.020	.002	.019	.001	.065	.056	.081	.080	.070	.068
2.5	.024	.004	.028	.001	.015	.001	.051	.038	.061	.060	.075	.073
3.0	.018	.002	.023	.004	.017	.001	.039	.024	.039	.038	.054	.052
Fractional area (% area/boundary)												
30/0	.017	.003	.032	.002	.016	.000	.021	.004	.014	.004	.019	.003
50/0	.027	.003	.020	.001	.023	.004	.023	.004	.027	.006	.033	.006
70/0	.024	.005	.024	.001	.019	.003	.030	.007	.017	.006	.027	.008
30/ – 1	.020	.001	.033	.003	.018	.001	.019	.003	.029	.008	.028	.004
50/ – 1	.023	.004	.035	.001	.022	.003	.025	.002	.031	.013	.031	.009
70/ – 1	.024	.005	.031	.004	.016	.003	.028	.007	.029	.011	.036	.008
30/ – 2	.025	.003	.033	.001	.018	.000	.021	.003	.036	.010	.027	.007
50/ – 2	.024	.006	.033	.001	.018	.001	.023	.003	.039	.015	.030	.010
70/ – 2	.023	.005	.029	.003	.014	.001	.026	.005	.029	.009	.033	.010

nonconcussed athletic control group were considered). The grand average ERP of the data set is depicted in Figure 5. The data set consisted of 18 participants with at least 198 artifact-free trials per participant (average 323 trials). The baseline period was 100 ms prior to stimulus onset, and a recording epoch lasted until 900 ms after stimulus onset. The sampling rate was 256 Hz, and the data were low-pass filtered at 67 Hz and baseline corrected. Trials with eyeblinks (VEOG > 80 μV), large horizontal eye movements (HEOG > 35 μV), and within-trial deviations (i.e., difference between the maximum and minimum voltage values in an epoch) exceeding 80 μV at Pz were rejected.

Compared to the data sets for the N1, the mean peak amplitude of the P3 is larger than the mean peak amplitude of the visual N1, but similar to the mean peak amplitude of the auditory N1. Inspection of each participant's average waveform revealed more variability (higher *SD*) of the P3 peak amplitudes than of both the visual and auditory N1 peak amplitudes.

The procedures for determining P3 onset latency were applied in the time window 200–900 ms after stimulus onset. To determine latency we used the peak latency technique, the absolute criterion technique with parameters 1.0, 2.0, . . ., 8.0 μV; the relative criterion technique with parameters 10%, 30%, 50%, 70%, or 90% of the peak amplitude; and the baseline deviation technique with parameters 2.0, 2.5, or 3.0 standard deviations. For the fractional area technique, parameters were 30%, 50%, or 70% of the area above the boundary combined with boundary values set to 0.0, 2.0, or 5.0 μV.

As before, linear interpolation was used to determine onsets with all techniques other than the peak latency. If the criterion value was not reached during the time window, the end of the time window (i.e., 900 ms) was taken as the latency. If the criterion value had already been reached at the beginning of the time window, the starting time of the window (i.e., 200 ms) was taken as the onset latency.

In the following, we present simulations indicating how accurately the P3 latency differences were estimated (Table 6) and simulations evaluating statistical power (Tables 7 and 8). For reasons of brevity we do not present between-subjects comparisons. Instead we present further simulations to evaluate the

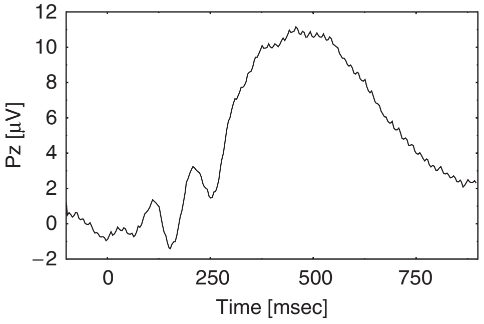

Figure 5. Grand average of the data set that was used for the P3 simulation: visually evoked P3 at Pz electrode.

Table 6. *Mean (M) and Standard Deviation (SD) in Milliseconds of the Estimated Differences (D) for P3 at Electrode Pz*[a]

	Visual data set			
	Single participants		Jackknife	
Criterion	*M*	*SD*	*M*	*SD*
Peak	14.86	24.88	14.97	27.96
Absolute criterion (μV)				
1.0	5.69	10.82	2.03	4.92
2.0	9.13	20.22	6.96	14.15
3.0	10.09	32.92	12.32	20.42
4.0	11.75	35.01	16.07	17.94
5.0	8.78	38.14	16.13	9.71
6.0	6.38	83.52	15.67	7.22
7.0	6.34	100.46	15.75	9.54
8.0	4.63	151.20	16.19	16.17
Relative criterion (% maximum amplitude)				
10	8.97	10.12	4.13	7.02
30	13.60	8.60	20.55	21.06
50	15.92	8.94	17.00	4.65
70	16.46	9.37	17.40	9.10
90	15.70	18.49	16.54	23.04
Baseline deviation (number of noise *SD*)				
2.0	7.10	34.47	5.12	11.97
2.5	6.04	34.50	7.01	17.99
3.0	6.47	30.47	8.18	23.39
Fractional area (% area/boundary)				
30/0	11.89	5.35	12.76	5.14
50/0	12.06	7.17	12.90	6.66
70/0	11.51	9.74	12.16	9.70
30/2	13.10	7.29	15.00	5.68
50/2	12.96	9.77	15.11	7.63
70/2	12.62	11.20	14.91	11.20
30/5	12.74	15.74	15.57	6.23
50/5	13.20	18.55	15.43	7.31
70/5	12.62	21.92	15.64	9.14

[a]The true mean difference was 15.625 ms.

fractional area technique depending on the size of the time window (Table 9), because it is somewhat problematic to determine the appropriate time window with a broad component like the P3.

The tables list the results in the same order used previously. To keep the results description as short as possible, we do not discuss each procedure. Instead we describe the preferable methods that perform most satisfactorily in terms of effect size estimation and power.

The section "Discussion of P3 Results" summarizes the main findings. Thus, readers not interested in result details may continue there.

Estimation of 15.625-ms Effects

Table 6 shows the results of simulations evaluating how accurately the single-participant and the jackknife approaches estimate P3 latencies when using different scoring techniques. In these simulations, $n = 12$ participants were chosen randomly without replacement from the available data pool of 18 participants. For each participant, 70 experimental and 70 control trials were also chosen randomly without replacement. Experimental trials were shifted exactly 15.625 ms. The simulation included 1000 experiments to estimate the differences (D) for the P3 latency in the experimental versus control conditions. Means and standard deviations (SD) of the estimated differences (D) are listed in Table 6. A scoring procedure is better the closer the mean to the true shift (15.625 ms) and the smaller the *SD*. For the single-participant approach, most of the techniques clearly underestimated the true difference. Only the peak latency technique and the relative criterion technique with parameters 50% or higher produced estimates that were approximately accurate on average, but all of these had much larger *SD* than the jackknife approach. Considering both mean and *SD*, the best estimates of the difference resulted from the jackknife approach combined with either the absolute criterion technique with parameters 5.0–7.0 μV, the relative criterion technique with parameters 50% and 70%, or the fractional area technique with parameters 30% and 50% of the area and a boundary of 2.0 or 5.0 μV.

Power to Detect 15.625-ms Effects

To estimate the power to detect the latency shift we ran simulations with number of participants, $n = 8$ and 12, and with number of trials, $t = 50$, 70, and 90. Significance level was set to $p = .05$ for each two-tailed t test. Again, each simulation included 1000 experiments. The proportions yielding significant t tests are listed in Table 7 as estimates of the power to detect the 15.625-ms shift.

First of all, the power to detect this latency shift is generally rather poor. Tolerable power estimates were only observed for $n = 12$ participants and $t = 70$ and 90 trials for the jackknife approach combined with the absolute criterion technique with the parameter 5.0 μV, the relative criterion technique with the parameter 50%, and the fractional area technique with parameters 30% and 50% with each boundary. In contrast to the N1 simulations, the relative criterion technique with the parameter 50% produced larger power and thus appears superior to the fractional area technique when combined with the jackknife approach.

The overall power to detect P3 latency differences was rather low in these simulations, presumably because these simulations were run with a small true effect size for P3 latency shifts. P3 latency differences around 30 ms or more are often obtained (e.g. Leuthold, & Sommer, 1998; Magliero et al., 1984; Smulders, Kok, Kenemans, & Bashore, 1995). To give an estimate of how well the procedures work with such effect sizes, we ran an additional simulation with a latency difference of 31.25 ms. Table 8 presents means and *SD* for estimating the difference as well as power to detect the latency shift resulting from a simulation with $n = 12$ participants and $t = 50$ trials per condition.

Not surprisingly, performance was best with the same procedures that worked well in the previous simulations. Especially accurate estimates of mean latency differences combined with high power to detect the latency shift were observed for the jackknife approach combined either with the relative criterion technique with parameter 50% or with the fractional area technique with parameter 30%, regardless of the boundary. For these procedures, power values were satisfactorily high (over 90%). Thus, P3 latency shifts of 30 ms and higher can be expected to produce significant results with a data set of 12 participants and at least 50 trials per condition.

Fractional Area Technique for Smaller Time Windows

The fractional area technique determines latency depending on the area under the ERP curve. However, for broadly extended ERPs like the P3, there might be cases in which it is not possible to consider the whole time window of the ERP curve. For example, if participants often blinked after performing a

Table 7. *Power to Detect Latency Difference of 15.625 ms as a Function of Method and Criterion, Depending on Number of Trials and Number of Participants for the P3 at Electrode Pz*

	8 participants						12 participants					
	Single participants			Jackknife			Single participants			Jackknife		
Criterion	50 trials	70 trials	90 trials	50 trials	70 trials	90 trials	50 trials	70 trials	90 trials	50 trials	70 trials	90 trials
Peak	.063	.077	.088	.017	.039	.036	.099	.107	.110	.041	.049	.061
Absolute criterion (μV)												
1.0	.044	.055	.067	.006	.007	.006	.081	.094	.136	.000	.001	.001
2.0	.079	.088	.102	.018	.026	.033	.099	.138	.137	.005	.007	.015
3.0	.074	.107	.150	.086	.094	.140	.107	.147	.179	.090	.109	.121
4.0	.089	.132	.159	.176	.231	.332	.104	.157	.166	.380	.443	.490
5.0	.094	.115	.173	.253	.325	.412	.101	.134	.159	.572	.651	.740
6.0	.114	.131	.156	.201	.288	.353	.102	.123	.152	.364	.447	.490
7.0	.072	.069	.084	.131	.159	.187	.064	.061	.061	.135	.160	.193
8.0	.030	.034	.030	.057	.070	.084	.041	.023	.011	.051	.067	.080
Relative criterion (% maximum amplitude)												
10	.048	.068	.076	.007	.005	.010	.117	.133	.147	.000	.001	.002
30	.267	.383	.470	.180	.202	.279	.377	.458	.540	.217	.247	.259
50	.300	.358	.432	.376	.487	.563	.388	.446	.489	.658	.751	.815
70	.258	.370	.462	.074	.128	.156	.377	.477	.565	.105	.146	.151
90	.107	.161	.190	.026	.037	.027	.137	.181	.214	.042	.031	.034
Baseline deviation (number of noise *SD*)												
2.0	.059	.066	.074	.009	.010	.010	.080	.104	.089	.006	.001	.000
2.5	.048	.054	.063	.012	.012	.012	.071	.095	.093	.010	.002	.001
3.0	.045	.058	.074	.014	.018	.020	.057	.099	.103	.019	.004	.003
Fractional area (% area/boundary)												
30/0	.311	.401	.473	.312	.404	.461	.479	.583	.648	.456	.594	.669
50/0	.194	.233	.284	.177	.262	.318	.276	.350	.408	.309	.378	.456
70/0	.127	.163	.180	.086	.123	.150	.161	.197	.247	.141	.172	.215
30/2	.281	.370	.448	.321	.419	.488	.424	.488	.562	.497	.617	.694
50/2	.154	.224	.266	.186	.267	.332	.254	.286	.354	.305	.382	.470
70/2	.116	.166	.187	.093	.141	.179	.177	.199	.241	.168	.195	.250
30/5	.143	.182	.243	.276	.375	.447	.182	.203	.220	.444	.548	.644
50/5	.095	.118	.182	.185	.266	.333	.133	.164	.172	.310	.398	.511
70/5	.074	.074	.115	.115	.184	.218	.089	.112	.105	.216	.261	.371

Table 8. *Mean (M) and Standard Deviation (SD) of Estimated Differences (D) and the Power ($1-\beta$) to Detect Latency Differences of 31.25 ms for the P3 at Electrode Pz*

	Single participants			Jackknife		
Criterion	*M*	*SD*	$1-\beta$	*M*	*SD*	$1-\beta$
Peak	28.96	25.80	.225	30.45	28.91	.082
Absolute criterion (μV)						
1.0	14.75	12.41	.315	15.47	8.92	.145
2.0	18.24	21.49	.332	23.32	16.38	.180
3.0	21.46	30.63	.320	30.61	23.24	.198
4.0	23.20	35.03	.320	32.50	20.25	.463
5.0	24.57	38.62	.299	31.26	10.67	.809
6.0	26.10	49.47	.302	31.25	7.91	.740
7.0	28.96	96.13	.201	31.59	10.76	.468
8.0	35.50	149.99	.090	30.38	16.24	.326
Relative Criterion (% maximum amplitude)						
10	22.96	12.31	.456	21.79	13.37	.215
30	31.26	11.63	.751	42.91	23.67	.308
50	33.40	11.02	.855	33.19	5.56	.943
70	33.34	12.95	.767	33.74	10.57	.455
90	31.49	21.86	.397	32.22	27.68	.091
Baseline deviation (number of noise *SD*)						
2.0	16.28	32.78	.230	19.03	15.60	.098
2.5	13.03	34.48	.177	20.55	21.55	.083
3.0	13.82	36.20	.136	21.47	27.23	.066
Fractional area (% area/boundary)						
30/0	25.16	6.66	.910	26.74	6.12	.962
50/0	24.33	8.66	.701	26.27	7.76	.850
70/0	22.88	11.58	.466	24.53	11.02	.471
30/2	26.77	8.54	.802	29.97	6.78	.958
50/2	26.25	10.89	.604	29.88	8.88	.815
70/2	25.00	12.67	.478	29.27	12.78	.494
30/5	25.73	17.64	.481	30.74	7.52	.911
50/5	25.61	20.36	.367	30.79	8.78	.830
70/5	24.30	23.37	.268	30.68	10.74	.676

Table 9. *Mean (M) and Standard Deviation (SD) of Estimated Differences (D) and the Power (1 − β) to Detect Latency Differences of 31.625 ms with the Fractional Area Technique Depending on the Size of the Time Window in Which the P3 is Scored at Electrode Pz*

	200–600 ms						200–900 ms					
	Single participants			Jackknife			Single participants			Jackknife		
Criterion	*M*	*SD*	$1-\beta$	*M*	*SD*	$1-\beta$	*M*	*SD*	$1-\beta$	*M*	*SD*	$1-\beta$
Fractional area (% area/boundary)												
30/0	19.43	4.32	.954	20.31	3.80	.992	25.29	5.56	.972	26.89	5.25	.992
50/0	16.80	4.49	.900	16.27	3.44	.960	24.68	7.64	.817	26.65	6.68	.944
70/0	13.18	3.85	.875	12.11	2.65	.971	23.27	9.79	.610	24.85	9.37	.647
30/2	21.30	6.55	.854	22.26	4.03	.990	27.22	7.68	.869	30.36	5.76	.991
50/2	18.46	7.09	.757	19.04	3.69	.977	26.66	9.87	.721	30.57	7.66	.929
70/2	14.74	7.46	.614	13.77	2.95	.964	25.37	11.37	.583	29.97	11.00	.658
30/5	20.53	13.65	.495	24.33	5.12	.944	25.54	15.66	.507	31.50	6.32	.967
50/5	19.03	14.51	.468	21.25	4.79	.938	26.11	17.99	.422	31.84	7.42	.946
70/5	15.21	16.36	.263	15.99	4.35	.779	24.98	20.85	.304	31.59	9.12	.813

response, many experimental trials would have to be excluded from the analysis when considering time windows of 900 ms or longer. Conversely, more experimental trials could be included in the analysis if the P3 time window could be restricted. Of course restricting the time window to such a degree that the decreasing part of the ERP component is no longer completely included has no impact on the latency criterion techniques referring to the increasing part of the curve, that is, to absolute, relative, baseline, and peak criterion techniques. However, the fractional area technique would be sensitive to such a restriction because this technique considers the whole area under the curve in the given time window to determine latency. To determine how much the fractional area technique is influenced by the size of the time window, we ran a simulation with $n = 12$ participants, $t = 70$ trials (for experimental and control conditions), and a latency shift of 31.625 ms. The simulation included 1000 experiments. In each experiment P3 latency was searched both in a restricted time window from 200 to 600 ms and in the usual time window from 200 to 900 ms with the fractional area technique with the parameters 30%, 50%, and 70% combined with boundaries of 0.0, 2.0, and 5.0 μV.

Means and *SD* for the obtained difference as well as power to detect the latency shift for both time windows are shown in Table 9. When the latency is searched in the restricted time window ranging from 200 to 600 ms, the size of the latency shift is substantially underestimated relative to the analysis including the larger time window. Nevertheless, the power to detect the shift remained high (in some cases it was even higher than for the long time window). Thus, reducing the time window may be a reasonable strategy if this enables the researcher to include many more experimental trials in the analysis, especially if detecting the presence or absence of a shift is more important than estimating the true size of the shift.

Type I Error

To estimate Type I error proportions for the P3 simulations we applied a simulation protocol identical to that of the simulations examining power for P3 latency differences, except that the experimental trials were not shifted along the time axis. Simulations were run for $n = 12$ participants and $t = 50$, 70, and 90 trials. They revealed that Type I error was generally low for each simulation procedure. As was true for the N1, then, inflation of Type I error rate is not a concern with any of these analysis procedures.

Discussion of P3 Results

The simulations regarding P3 again show that the standard procedure for evaluating ERP latencies, that is, single-participant comparisons of peak latencies, is not the most efficient one. Instead, the jackknife approach combined with either the relative criterion technique with parameter 50% or with the fractional area technique with parameter 30% of the area and boundary of 2.0 μV or higher turned out to be most useful.

Further, these simulations revealed that P3 latency shifts tend to be more difficult to detect than N1 latency shifts, even if the true effect sizes are larger. Therefore, we consider it important to be cautious with null results, especially in settings for which one would expect rather small latency differences or when the numbers of participants and experimental trials are small. If possible, we recommend using a data set of at least 12 participants (of course, the more the better) and at least 70 trials for each experimental condition.

Interestingly, comparing the simulations involving N1 and P3 latency shifts shows that the advantage of the jackknife approach combined with either the relative criterion or the fractional area technique becomes larger the more noisy the data. Thus, we generally recommend the use of these methods, especially if the signal-to-noise ratio in the data is low.

However, the simulations also reveal that the jackknife approach is not necessarily advantageous for every data set or for every latency onset technique. In particular, it appears that the jackknife approach should not be combined with the peak latency technique, because this combination produces rather low power. Thus, if in some instances the peak latency technique seems especially appropriate, the jackknife approach should be avoided.

Simulations for the N2pc Component

In this third part of the article, the same methods that were tested for the N1 and P3 components were applied to the N2pc component. The N2pc is a lateralized ERP component that is maximal at posterior electrode sites contralateral to an attended item. It is isolated by subtracting activity at ipsilateral electrode sites from the corresponding activity at contralateral electrode sites (e.g., PO7/PO8). The N2pc typically starts about 180 ms after target onset and lasts about 100 ms. Luck and colleagues, who were the first to study this component meticulously in visual search tasks, suggested that the N2pc reflected distractor suppression processes (Luck & Hillyard, 1994; Luck, Girelli, McDermott, & Ford, 1997). Others, who have used bilateral displays with only one distractor, have argued that the N2pc reflected target enhancement processes (e.g., Eimer, 1996). Nonetheless, even if there is still an ongoing debate on the specific processes that underlie the N2pc, it is widely accepted as a valid index of covert visual–spatial attention in light of several results reviewed by Woodman and Luck (2003), and it has been widely used in the study of visual–spatial attention (e.g., Brisson & Jolicœur, 2007a, 2007b, 2007c; Dell'Acqua, Sessa, Jolicœur, & Robitaille, 2006; Eimer & Mazza, 2005; Girelli & Luck, 1997; Jolicœur, Sessa, Dell'Acqua, & Robitaille, 2006a, 2006b; Wascher, 2005; Woodman & Luck 2003).

The N2pc differs from N1 and P3 because it is measured from a difference wave, which tends to increase variability. Furthermore, the signal is generally smaller for the N2pc component than for the N1 and P3 components. So these two characteristics would tend to make it more difficult to detect latency differences. But, on the other hand, the N2pc peak is better defined than for the P3 component, a characteristic most likely facilitating the detection of latency differences.

General Simulation Protocol

The general protocol for simulations examining N2pc latency differences was similar to the protocol for the simulations of N1 and P3 latency differences.

For the simulations, we took data from an experiment with visual stimulation of both the left and right visual hemifield with attention deployed to a target stimulus in one hemifield (for a detailed description of the experiment, see De Beaumont et al., 2007; for the simulations, we took only the data of the single-concussion and multiconcussion groups). A difference wave was computed for each participant by subtracting the ipsilateral waveform (left-sided electrode with left visual field target and right-sided electrode with right visual field target) from the contralateral waveform (left-sided electrode with right visual field target and right-sided electrode with left visual field target) at the PO7 and PO8 electrodes, and N2pc latencies were obtained from these difference waves. The grand average of the data set is depicted in Figure 6. The data set consisted of 30 participants with at least 149 artifact-free trials per participant per condition (i.e., left or right target; average 227 trials). The baseline period was 200 ms prior to stimulus onset, and a recording epoch lasted until 500 ms after stimulus onset. The sampling rate was 256 Hz, and the data were low-pass filtered at 67 Hz and baseline corrected. Trials with eyeblinks (VEOG > 80 μV), large horizontal eye movements (HEOG > 35 μV), and within-trial deviations (i.e., difference between the maximum and minimum voltage values in an epoch) exceeding 80 μV at PO7 and/or PO8 were rejected.

The peak amplitude of the N2pc is less than the peak N1 and P3 amplitudes. Inspection of individual participants' waveforms revealed that the variability of peak amplitude is smaller than for the other data sets.

The procedures for determining N2pc onset latency differences were applied in the time window 0–350 ms after stimulus onset. To determine latency we used the peak latency technique, the absolute criterion technique with parameters – 0.25, – 0.50 . . . , – 2.0 μV; the relative criterion technique with parameters 10%, 30%, 50%, 70%, or 90% of the peak amplitude; and the baseline deviation technique with parameters 2.0, 2.5, or 3.0 standard deviations. For the fractional area technique, parameters were 30%, 50%, or 70% of the area below the boundary combined with boundary values set to 0.0, – 0.1, or – 0.5 μV.

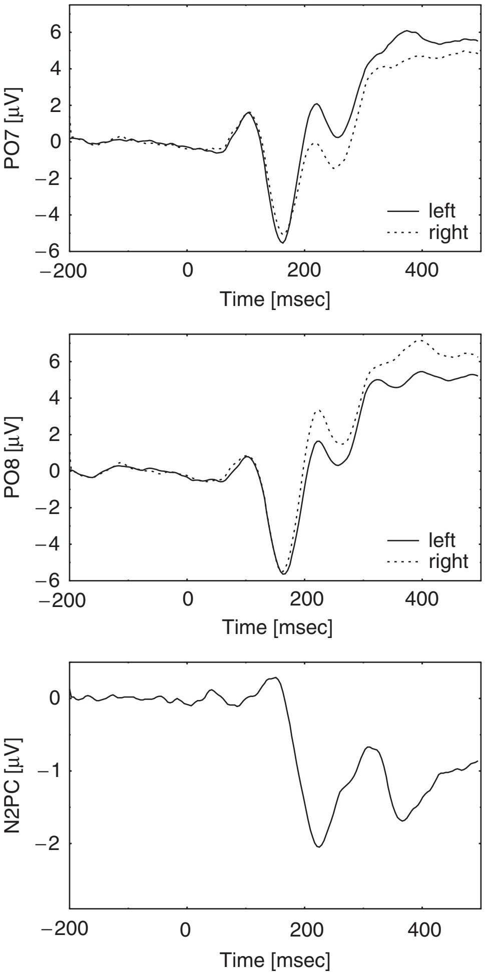

Figure 6. Grand average of the data set that was used for the N2pc simulation. Upper and middle panels: visually evoked N2 at PO7 and PO8 electrodes depending on target location. Lower panel: resulting N2pc difference wave.

As before, linear interpolation was used to determine onsets with all techniques other than the peak amplitude. If a parameter value had already been reached at the beginning of the time window or was not reached during the time window, the starting time of the window (i.e., 0 ms) or the end of the window (i.e., 350 ms) was taken as latency.

We decided to shift the EEG data of the experimental trials 31.25 ms, which lies in the range of observed effect sizes in a representative study (Wascher, 2005). In the following, we present simulations indicating how accurately these N2pc latency differences were estimated (see Table 10), simulations evaluating statistical power (see Table 11), and we briefly refer to simulations estimating Type I error probability. Additionally, we included simulations evaluating between–subjects comparisons (see Table 12) because we were not sure whether between-subjects and within-subject comparisons for difference waves, like the N2pc, reveal similar results (as obtained for between-subjects and within-subject comparisons for the N1).

The tables list the results in the same order as before. The description of the results is kept rather short. We just describe

Table 10. *Mean (M) and Standard Deviation (SD) in Milliseconds of the Estimated Differences (D) for the N2pc at Electrodes PO7/PO8*[a]

	Single subjects		Jackknife based	
Criterion	*M*	*SD*	*M*	*SD*
Peak	13.43	22.95	29.31	15.59
Absolute criterion (μV)				
– 0.25	– 2.52	22.85	7.46	89.71
– 0.50	0.87	30.75	28.59	64.91
– 0.75	8.20	35.23	30.98	19.96
– 1.00	14.66	34.96	31.45	8.88
– 1.25	18.96	32.70	31.49	13.86
– 1.50	20.52	29.91	30.43	29.00
– 1.75	21.14	28.10	27.96	43.12
– 2.00	18.59	27.69	20.35	50.94
Relative criterion (% maximum amplitude)				
10	1.98	23.41	2.96	82.68
30	5.11	29.83	27.42	47.13
50	16.39	27.36	30.77	10.53
70	21.04	24.15	30.97	6.12
90	15.53	23.80	30.69	9.00
Baseline deviation (number of noise *SD*)				
2.0	8.70	44.67	17.73	94.10
2.5	11.35	39.98	23.88	73.68
3.0	13.63	35.25	27.01	56.27
Fractional area (% area/boundary)				
30/0	14.03	17.83	25.75	6.78
50/0	17.32	13.71	22.78	5.90
70/0	14.62	11.34	15.40	9.10
30/ – 0.1	15.57	18.03	26.82	5.65
50/ – 0.1	17.76	14.32	23.41	5.83
70/ – 0.1	14.76	12.11	16.17	9.75
30/ – 0.5	20.11	19.19	28.81	4.41
50/ – 0.5	19.27	17.47	25.89	6.61
70/ – 0.5	15.55	16.72	19.86	12.52

[a]The true mean difference was 31.25 ms.

the preferable methods that perform satisfyingly in terms of effect estimation and power to detect them.

The section "Discussion of N2pc Results" summarizes the main findings. Readers not interested in result details are referred directly there.

Estimation of 31.25-ms Effects

Table 10 shows the results of a simulation evaluating how accurately the single participant and the jackknife approach estimate N2pc latencies when we use different scoring techniques. In the simulation, $n = 12$ participants were chosen randomly without replacement from the available data pool of 30 participants. For each participant, 50 experimental and 50 control trials for each of the conditions left and right target presentation were also chosen randomly without replacement; that is, the total number of trials for the experimental and control conditions was 100. Experimental trials were shifted exactly 31.25 ms. The simulation includes 1000 single experiments to estimate the differences (*D*) for the N2pc latency in the experimental versus control conditions. Means and standard deviations (*SD*) of the estimated difference (*D*) are listed in Table 10. A scoring procedure is better the closer the mean is to the true shift (31.25 ms) and the smaller the *SD*. For the single–participant approach, most of the techniques clearly underestimated the true difference and had relatively large *SD*. Although the jackknife-based approach also generally underestimated the true difference, this approach provided the best estimates of the difference considering both mean and *SD*, when it was combined with the absolute criterion technique with parameters – 0.75 to – 1.25 μV, the relative criterion technique with parameters 50% and higher, or the fractional area technique with parameters 30% and 50% of the area and a negative boundary.

Power to Detect 31.25–ms Effects

To estimate the power to detect the latency shift we ran simulations with number of participants, $n = 8$, 12, and 20, and with number of trials, $t = 60$, 100, and 140 (please note that this is the sum of the number of left-target and right-target trials; for each simulation equal numbers of left-target and right-target trials were chosen). Significance level was set to $p = .05$ for each two-tailed *t* test. Again, each simulation included 1000 experiments. The proportions yielding significant *t* tests are listed in Table 11 as an estimate of power to detect the 31.25 ms shift.

As expected, power increased with increasing numbers of participants and trials, reflecting the usual impact of these variables on statistical power for all procedures. Power was quite low for the single–participants approach, independently of the scoring technique and parameter used. For the jackknife approach, high power was obtained with the absolute criterion technique with the parameters – 0.75 to – 1.25 μV, with the relative criterion technique with the parameters 50% and higher, and with the fractional area technique with parameters 30% and 50% with each boundary. As was the case for the N1 simulations, the fractional area technique when combined with the jackknife approach yielded the largest power, followed closely by the combination of the jackknife approach and the relative criterion technique with the parameters 50% and higher.

Type I Error

To estimate Type I error proportions, the same simulation protocol as before was applied except that experimental trials were not shifted along the time axis. Simulations with number of participants $n = 12$ and number of trials $t = 60$, 100, and 140 yielded satisfyingly low Type I error proportions for all scoring procedures. Thus, as for the N1 and P3 simulations, Type I error is not a concern with any of the procedures.

Between–Subjects Comparison

For a more complete evaluation of the different procedures for estimating N2pc latency differences, we also ran simulations examining between-subjects comparisons. To implement a between-subjects comparison, the simulation protocol was changed slightly. In each simulation step, 16 participants were chosen randomly. Half of them were assigned to the control group, and the other half were assigned to the experimental group. From the data pool of each participant, 100 trials were chosen randomly (50 left- and 50 right-target trials). Then the entire waveforms for all of the trials of the experimental group were shifted 31.25 ms in time. After this shift, the data for the experimental and control groups were analyzed as usual with each scoring procedure. The obtained latency values were compared with between-subjects *t* tests with a two-tailed significance level of $p = .05$. The *t* values for the jackknife approach were adjusted according to formula 1 with $n =$ number of participants per group (i.e., $n = 8$).

The main simulation results, including estimates of Type I error proportions, are shown in Table 12. The single-participant

Table 11. *Power to detect Latency Differences of 31.25 ms as a Function of Method and Criterion, Depending on Number of trials (Sum of the Number of Left-Target and Right-Target Trials) and Number of Participants for the N2pc in Data Set A at Electrodes PO7/PO8*

	8 participants						12 participants						20 participants					
	Single participants			Jackknife			Single participants			Jackknife			Single participants			Jackknife		
Criterion	60 trials	100 trials	140 trials	60 trials	100 trials	140 trials	60 trials	100 trials	140 trials	60 trials	100 trials	140 trials	60 trials	100 trials	140 trials	60 trials	100 trials	140 trials
Peak	.067	.102	.104	.315	.391	.421	.075	.108	.132	.445	.536	.594	.098	.135	.146	.614	.782	.847
Absolute criterion (μV)																		
− 0.25	.022	.016	.009	.040	.052	.044	.013	.013	.014	.053	.061	.082	.014	.014	.013	.080	.109	.126
− 0.50	.015	.027	.019	.068	.190	.252	.018	.015	.016	.141	.368	.618	.020	.026	.019	.448	.783	.951
− 0.75	.018	.038	.038	.189	.501	.691	.030	.044	.049	.475	.817	.932	.030	.043	.050	.885	.985	.997
− 1.00	.033	.055	.063	.344	.649	.787	.036	.058	.089	.668	.854	.930	.053	.072	.099	.898	.983	.991
− 1.25	.047	.086	.124	.382	.568	.675	.059	.095	.136	.622	.783	.873	.063	.123	.165	.862	.964	.989
− 1.50	.057	.124	.147	.292	.414	.472	.066	.106	.177	.493	.610	.667	.075	.162	.188	.734	.869	.917
− 1.75	.057	.128	.136	.201	.256	.283	.078	.129	.188	.302	.366	.383	.113	.174	.180	.431	.563	.579
− 2.00	.064	.110	.128	.122	.135	.138	.076	.121	.168	.151	.154	.159	.105	.179	.163	.176	.182	.167
Relative criterion (% maximum amplitude)																		
10	.021	.013	.008	.030	.036	.034	.021	.019	.019	.038	.048	.059	.026	.030	.018	.054	.065	.068
30	.031	.032	.046	.134	.386	.507	.035	.029	.043	.349	.653	.802	.039	.042	.046	.715	.938	.986
50	.065	.100	.159	.507	.750	.843	.077	.100	.154	.799	.923	.977	.058	.100	.166	.966	.995	.999
70	.106	.180	.241	.603	.754	.835	.118	.182	.227	.812	.939	.961	.117	.195	.258	.956	.996	1.00
90	.089	.137	.127	.451	.572	.643	.093	.126	.143	.675	.779	.854	.100	.141	.161	.864	.958	.983
Baseline deviation (number of noise *SD*)																		
2.0	.035	.041	.039	.081	.123	.106	.036	.052	.039	.123	.138	.123	.034	.047	.036	.169	.150	.115
2.5	.043	.052	.055	.093	.197	.201	.047	.050	.050	.186	.236	.252	.044	.069	.048	.296	.291	.219
3.0	.043	.054	.082	.114	.261	.304	.047	.068	.086	.237	.365	.398	.066	.075	.080	.408	.432	.416
Fractional area (% area/boundary)																		
30/0	.088	.153	.216	.443	.662	.769	.086	.159	.215	.666	.857	.946	.098	.182	.232	.895	.979	.997
50/0	.207	.360	.492	.510	.687	.791	.214	.379	.458	.717	.874	.939	.261	.411	.519	.927	.988	.998
70/0	.219	.328	.447	.184	.247	.292	.228	.394	.435	.207	.294	.320	.300	.459	.569	.274	.356	.435
30/ − 0.1	.106	.193	.258	.541	.770	.871	.096	.182	.254	.778	.931	.980	.110	.219	.282	.965	.994	.999
50/ − 0.1	.222	.356	.499	.529	.710	.807	.219	.393	.479	.738	.896	.950	.258	.412	.523	.952	.989	.999
70/ − 0.1	.207	.325	.437	.181	.247	.287	.212	.376	.415	.205	.280	.308	.289	.432	.537	.267	.336	.393
30/ − 0.5	.161	.274	.326	.776	.919	.961	.147	.255	.329	.948	.994	1.00	.156	.295	.339	1.00	1.00	1.00
50/ − 0.5	.192	.321	.435	.535	.675	.781	.211	.345	.463	.720	.876	.913	.230	.344	.451	.913	.977	.997
70/ − 0.5	.156	.259	.336	.219	.254	.314	.154	.263	.307	.238	.307	.351	.213	.308	.388	.294	.334	.375

Table 12. *Mean (M) and Standard Deviation (SD) of Estimated Differences (D) and the Power (1 – β) to Detect Latency Differences of 31.25 ms and Type I Error (α) in a Between-Subjects Design with 16 Participants and 100 Trials (50 Left-Target and 50 Right-Target Trials) for the N2pc at Electrodes PO7/PO8*

	Single participants				Jackknife			
Criterion	*M*	*SD*	1 – β	α	*M*	*SD*	1 – β	α
Peak	13.93	35.26	.064	.014	25.46	25.45	.322	.036
Absolute criterion (μV)								
– 0.25	– 3.65	29.28	.008	.013	– 3.03	81.87	.045	.044
– 0.50	– 0.61	40.38	.018	.026	17.48	86.50	.081	.040
– 0.75	5.30	47.98	.023	.018	28.46	46.21	.305	.027
– 1.00	13.02	48.44	.038	.012	29.97	25.04	.376	.030
– 1.25	18.42	46.73	.053	.023	29.94	29.66	.288	.026
– 1.50	20.74	45.55	.070	.021	28.18	47.45	.160	.035
– 1.75	19.87	44.08	.052	.021	25.01	68.32	.099	.057
– 2.00	19.61	42.72	.053	.020	21.27	81.31	.108	.101
Relative criterion (% maximum amplitude)								
10	2.22	30.20	.012	.016	– 0.97	77.87	.044	.047
30	5.86	43.43	.030	.025	19.49	73.77	.227	.023
50	14.96	42.15	.058	.023	29.43	25.70	.580	.036
70	18.53	37.69	.105	.020	30.79	14.20	.636	.036
90	15.15	35.86	.088	.013	29.23	18.26	.514	.033
Baseline deviation (number of noise *SD*)								
2.0	9.03	53.18	.040	.030	17.00	85.55	.076	.030
2.5	9.50	50.81	.046	.026	22.25	67.14	.122	.023
3.0	13.09	45.30	.048	.023	27.31	54.16	.163	.011
Fractional area (% area/boundary)								
30/0	13.97	29.04	.090	.018	24.56	12.70	.460	.013
50/0	17.86	21.92	.174	.006	22.60	12.94	.317	.026
70/0	15.63	19.03	.131	.015	15.53	18.23	.110	.018
30/ – 0.1	15.86	29.26	.098	.015	25.87	11.67	.530	.015
50/ – 0.1	18.39	22.70	.173	.008	23.29	13.34	.316	.028
70/ – 0.1	15.85	19.98	.125	.013	16.25	19.14	.115	.024
30/ – 0.5	20.13	30.69	.126	.010	28.43	10.88	.641	.019
50/ – 0.5	20.49	26.34	.162	.007	25.72	14.94	.305	.035
70/ – 0.5	16.43	25.44	.095	.010	19.28	22.66	.146	.040

approach does not reveal accurate estimates of the mean difference for any scoring technique. Likewise, the power to detect latency shifts is very low for this approach regardless of the latency estimation technique. Accurate estimates of the mean differences combined with satisfactorily low *SD* for these estimates and relatively higher power values were obtained for the jackknife approach combined with the relative criterion technique with parameters of 50% and higher and the fractional area technique with parameters 30% of the area and any boundary. It should be emphasized, though, that power to detect latency shifts was generally rather low for between-subjects comparisons, indicating the need for caution in interpreting null results in such comparisons. Type I error rate was not increased for any scoring procedure.

Discussion of N2pc Results

The standard procedure for evaluating ERP latencies, that is, single–participant comparisons of peak latencies, turned out to be rather ineffective when searching latency differences for the N2pc. Instead, the jackknife approach combined with either the relative criterion technique with parameter 50% or with the fractional area technique with parameter 30% or 50% of the area and any boundary was clearly to be preferred.

As was true for the P3, the N2pc simulations revealed that latency shifts are more difficult to detect than N1 latency shifts even if the expected effect sizes are larger. Therefore, researchers have to be cautious with null results, especially in settings when one would expect rather small latency differences or when the numbers of participants and experimental trials are small.

Furthermore, when we compare these simulations with the N1 and P3 results, the simulations confirm our previous assumption that the advantage of the jackknife approach combined with either the relative criterion or the fractional area technique becomes larger the more noisy the data. Especially for between-subjects comparisons, the jackknife approach combined with either the relative criterion or fractional area technique were the only reasonably effective methods.

Simulations for the Frequency-Related P3 Component

In this fourth and final section of the article, the same methods that were tested for the N1, P3, and N2pc components were applied to the frequency-related P3 component. The frequency-related P3 has also been studied using difference waves to isolate specific processes, as is done to isolate the N2pc, and this has been particularly useful for eliminating task overlap in dual-task studies. For example, several dual-task studies have taken advantage of the fact that the P3 amplitude is sensitive to the probability (frequency) of the class-defined stimulus category to isolate the P3 elicited by a second-task target from overlapping first-task target activity. This is done by subtracting the ERP elicited by a frequent second-task target category from the ERP elicited by an infrequent second-task target category for each task overlap condition. The onset latency of the resulting frequency-related P3 can then be taken as a relatively pure measure of the time required to perceive and categorize the second-task target in different overlapping conditions, devoid of Task 1 contamination (see Arnell, Helion, Hurdelbrink, & Pasieka, 2004; Dell'Acqua, Jolicœur, Vespignani, & Toffanin, 2005; Luck, 1998; Vogel & Luck, 2002).

Thus, as for the N2pc, the frequency-related P3 is measured from a difference wave. However, the frequency-related P3 is broader (i.e., more extended in time) and larger than the N2pc. Furthermore, the latency differences can be somewhat larger (e.g., more than 40 ms), especially in dual-task studies (see Arnell et al., 2004; Dell'Acqua et al., 2005; Luck, 1998; Vogel & Luck, 2002).

General Simulation Protocol

The general protocol for simulations examining frequency-related P3 latency differences was similar to the protocol for simulations of N1, P3, and N2pc latency differences. A difference wave was computed for each participant by subtracting the ERP waveform at the Pz electrode site for the frequent target category condition from the ERP waveform at the Pz electrode site for the infrequent target category condition, and frequency-related P3 latencies were obtained from these difference waves. We decided to shift the EEG data of the experimental trials 46.875 ms, which lies in the range of observed effect sizes (see Arnell et al., 2004; Dell'Acqua et al., 2005; Luck, 1998; Vogel & Luck, 2002).

For the simulations, we took data from an experiment with visual stimulation in which the P3 was recorded at the Pz electrode (for a detailed description of the experiment, see De Beaumont et al., 2007; only data of the nonconcussed athletic control group were considered). The grand average of the data set is depicted in Figure 7. The data set consisted of 18 participants with at least 198 artifact-free trials per participant in the frequent condition (average 323 trials) and at least 71 trials in the infrequent condition (average 108 trials). The baseline period was 100 ms prior to stimulus onset, and a recording epoch lasted until 900 ms after stimulus onset. The sampling rate was 256 Hz and the data were low-pass filtered at 67 Hz and baseline corrected. Trials with eyeblinks (VEOG > 80 μV), large horizontal eye movements (HEOG > 35 μV), and within-trial deviations (i.e., difference between the maximum and minimum voltage values in an epoch) exceeding 80 μV at Pz were rejected.

Compared to the N2pc, the peak amplitude of the frequency-related P3 is larger, but inspection of individual participants' different waveforms also revealed more variability.

The procedures for determining frequency-related P3 onset latency were applied in the time window 200–900 ms after stimulus onset. To determine latency we used the peak latency technique; the absolute criterion technique with parameters 0.5, 1.0, . . ., 4 μV; the relative criterion technique with parameters 10%, 30%, 50%, 70%, or 90% of the peak amplitude; and the baseline deviation technique with parameters 2.0, 2.5, or 3.0 standard deviations. For the fractional area technique the parameters were 30%, 50%, or 70% of the area above the boundary combined with boundary values set to 0.0, 0.5, or 1.0 μV.

As before, linear interpolation was used to determine onsets with all techniques other than the peak amplitude. If a parameter value had already been reached at the beginning of the time window or was not reached during the time window, the starting time of the window (i.e., 200 ms) or the end of the window (i.e., 900 ms) was taken as latency.

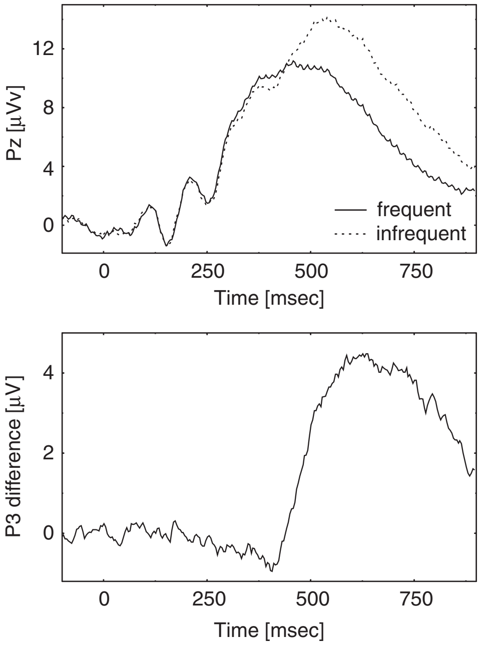

Figure 7. Grand average of the data set that was used for the frequency-related P3 simulation. Upper panel: visually evoked P3 at Pz electrode depending on stimulus frequency. Lower panel: resulting frequency-related P3 difference wave.

In the following, we present simulations indicating how accurately the frequency-related P3 latency differences were estimated and simulations evaluating statistical power and Type I error rate (all presented in Table 13). Of course the table is structured in the same way as before. For reasons of brevity we just describe the most preferable procedures and we do not present between-subjects comparisons.

Estimation and Power to Detect 46.875 ms Effects; Type I Error

Table 13 shows the results of simulations evaluating how accurately and with what power the single-participant and the jackknife approaches estimate frequency-related P3 latencies when we use different scoring techniques and the results of simulations estimating the Type I error to wrongly detect a nonexisting latency shift. In both simulations, the number of participants was $n = 18$. For each participant, 35 experimental trials and 35 control trials in the infrequent condition and 95 experimental trials and 95 control trials in the frequent condition were chosen randomly without replacement.[1] Experimental trials were shifted exactly 46.875 ms. The simulation includes 1000 single experiments to estimate the differences (D) for the frequency-related P3 latency in the experimental versus control conditions. Means and standard deviations (SD) of the estimated differences (D) are listed in Table 13. A scoring procedure is better the closer is the mean to the true shift (46.875 ms) and the smaller is the SD. For the single-participant approach, most of the techniques clearly underestimated the true difference and had

[1] In the original study, the ratio was 25:75 (35:115), which is slightly different from the 35:95 ratio chosen here. We wanted at least 35 trials in the infrequent condition, and the fact that there are fewer frequent trials (95 instead of 115) than the original ratio is not of concern, because it is the infrequent condition that drives the signal-to-noise ratio down.

Table 13. *Mean (M) and Standard Deviation (SD) of Estimated Differences (D), the Power (1 − β) to Detect Latency Differences of 46.875 ms, and Type I Error (α) for 18 Participants with 35 Trials in the Infrequent and 95 Trials in the Frequent Condition for Frequency-Related P3 Searched in the time window 200–900 ms, at Electrode Pz*

	Single participants				Jackknife			
Criterion	*M*	*SD*	1 − β	α	*M*	*SD*	1 − β	α
Peak	29.91	39.70	.125	.021	47.97	55.06	.083	.031
Absolute criterion (μV)								
0.5	− 8.34	39.09	.007	.025	− 30.63	170.03	.103	.078
1.0	− 7.77	46.46	.013	.026	33.25	70.92	.378	.011
1.5	− 7.95	54.61	.023	.026	46.84	22.16	.448	.015
2.0	− 5.58	60.07	.025	.028	47.36	17.32	.456	.017
2.5	2.09	61.23	.024	.015	47.26	19.02	.416	.008
3.0	8.69	60.81	.030	.016	47.42	28.34	.173	.017
3.5	15.70	55.55	.036	.015	48.14	46.33	.078	.006
4.0	23.08	62.83	.052	.015	43.26	91.57	.038	.003
Relative criterion (% maximum amplitude)								
10	− 2.61	39.50	.018	.029	− 50.13	166.85	.074	.077
30	1.23	42.87	.024	.028	44.95	24.24	.532	.017
50	13.28	42.88	.052	.027	46.91	13.26	.654	.012
70	24.52	41.50	.118	.019	47.94	25.46	.199	.025
90	29.82	40.23	.125	.021	48.51	33.93	.132	.020
Baseline deviation (number of noise *SD*)								
2.0	− 37.75	63.83	.007	.029	− 59.56	163.08	.066	.048
2.5	− 39.05	90.07	.006	.020	− 47.11	147.75	.077	.028
3.0	− 38.37	114.73	.007	.020	− 31.81	131.63	.104	.015
Fractional area (% area/boundary)								
30/0	22.96	28.08	.172	.016	39.83	14.30	.651	.011
50/0	24.99	28.01	.200	.012	35.50	14.15	.568	.011
70/0	22.77	28.18	.169	.011	31.01	12.85	.545	.008
30/0.5	24.68	30.98	.139	.013	42.34	13.33	.788	.013
50/0.5	25.97	31.53	.153	.012	38.24	15.20	.570	.010
70/0.5	24.18	31.99	.143	.012	34.16	14.81	.516	.009
30/1.0	26.13	33.97	.119	.016	43.91	14.20	.777	.012
50/1.0	26.93	34.90	.135	.017	40.67	17.19	.525	.011
70/1.0	24.99	35.31	.122	.016	37.43	17.66	.444	.014

relatively large *SD*.[2] The jackknife-based approach resulted in the best estimates of the difference, considering both mean and *SD*, when it was combined with either the peak latency technique, the absolute criterion technique with parameters 1.5 to 3.0 μV, the relative criterion technique with parameters 30% to 90%, or the fractional area technique with parameters 30% and 50% of the area and a boundary of 0.5 or 1.0 μV.

As was the case for the P3 component, power to detect the latency shift was generally rather poor, especially with the single-participant approach. Tolerable power estimates were only observed for the jackknife approach combined with the relative criterion technique with the parameter 30% and 50% and fractional area technique with 30% and 50% parameters combined with any boundary. Table 13 further shows that Type I error was within the nominal limits, as has already been shown for all the other components.

Discussion of Frequency-Related P3 Results

Simulations for the frequency-related P3 component revealed findings similar to those for the P3 and the N2pc. Again, the standard procedure for evaluating ERP latencies, that is, single-participant comparisons of peak latencies, turned out to be rather ineffective. Instead, the jackknife approach combined with either the relative criterion technique with parameter 30% and 50% or with the fractional area technique with parameter 30% and 50% of the area and any boundary turned out to be most useful.

Even if the expected effect sizes are larger for the frequency-related P3 component than for P3 or N2pc, simulations revealed that latency shifts are hard to detect. Thus, researchers must be very cautious with null results and should do everything possible to increase power, such as increase the number of participants and/or the number of trials.

Finally, the simulations for the frequency-related P3 component reinforce our previous conclusion that the advantage of the jackknife approach combined with either the relative criterion or fractional area technique becomes larger the more noisy the data.

General Discussion

This article reports simulations conducted to evaluate procedures for measuring changes in the onset latency of five different ERP components: the visual N1, the auditory N1, the P3, the N2pc (difference between event-related potentials at electrode sites contralateral and ipsilateral to an attended item), and the

[2] It may seem odd that we obtained consistently negative values for the single-subject approach combined with some techniques. Please note that this is the case if the estimate of the latency for the ERP is small (absolute criterion technique with parameters of 0.5 to 2.0, relative criterion technique with parameter 10%, and baseline deviation techniques). The frequency-related P3 component is characterizes by a small negative-going shift in the time window from 200 to 400 (see Figure 7). The estimated onset latency is searched in the time window 200–900 ms after stimulus onset. It is more probable that the criterion is already reached in the "noisy" time window 200–400 ms in the shifted data than in the original data, because for the shifted data also the small negative-going wave is shifted in time (and thus the small negativity that makes it less likely that the criterion is reached early starts later).

frequency-related P3 (difference between event-related potentials of infrequent and frequent target conditions). The evaluated measurement procedures were the single-participant approach and the jackknife approach combined with various specific techniques to estimate the onset latency of the component. These techniques included measurements based on peak latency, an absolute criterion, a relative criterion, baseline deviation, and fractional area. For each technique (except for peak latency) several parametric variations were evaluated.

The simulations revealed that the standard procedure for evaluating latency effects, that is, comparing the peak latencies for single participants, is definitely not the most efficient one for any of the components examined. If the signal-to-noise ratio is high, this procedure estimates and detects latency differences satisfactorily. However, as the signal-to-noise ratio drops, the power of this method to detect latency differences rapidly becomes worse.

In general, the jackknife approach combined with either the relative criterion technique or the fractional area technique turned out to be the most desirable methods. Thus, generally, we recommend that researchers test for component latency differences by combining the jackknife approach with either the relative criterion technique or else with the fractional area technique. When using the relative criterion technique, the 50% criterion appears to be a good choice in general. When using the fractional area technique, the optimal parameters appear to be a criterion of 30% or 50% of the area, with a slightly negative boundary for negative-going components and with a slightly positive boundary for positive-going components.

We must emphasize, however, that these recommendations are somewhat crude generalizations and that there are also cases for which these recommendations do not hold. For example, researchers always have to consider their data set carefully to judge whether it is sensible to take the area (and, if so, which fraction of the area) under the ERP component into account. Just consider the example in Figure 2 (right side). In this example, the fractional area technique with the parameter 50% of the area is not suitable for detecting differences in onset latency because for both the experimental and control groups 50% of the area of the curve is reached at the same point in time. Likewise, one can easily imagine data sets for which the fractional area technique would wrongly reveal latency differences, for example, if, the N1 has the same onset in two conditions but is closely followed or even partly overlapped by a component that varies in latency.[3] Such a differential overlap would lead to a reduction in total area of the N1 in one condition if the time window in which the latency criterion is searched still included the subsequent component. Consequently, the fractional area technique would most likely reveal an N1 latency difference. To avoid such a "false alarm," the time window should be restricted so that the overlapping component is excluded from the analyzed area. This most likely would not influence power to detect N1 latency differences (if there were any) as has been shown in the simulations regarding the P3 searched in restricted compared to more extended time windows (see Table 9).

In general, researchers should always first consider the visual appearance of the ERP waveforms in order to choose an appropriate procedure when determining the time window for the investigated component and indeed when deciding whether statistical tests of onset latency differences make any sense at all. We hope that the detailed results presented in the tables help researchers choose appropriate procedures when necessary. In this regard, we would like to mention that it seems unproblematic to analyze latency differences of two ERPs with several procedures (most likely with the jackknife approach combined with relative criterion and fractional area criterion and several different parameters) because Type I error level for each procedure is low (almost always lower than the estimated α level).

Finally, we want to point out again that null effects have to be treated very cautiously, especially when based on small sample sizes and/or few trials per conditions. Power to detect latency differences can be low, particularly for between-subjects comparisons. In addition to determining the most efficient method to detect latency differences, the reported simulations for the respective components may be useful as hints about how many participants and how many trials per condition are needed to obtain satisfactory power to detect latency differences in the range of the effect sizes that we have simulated. Thus, for N1 studies with expected effect sizes of 8 ms, power estimates are satisfactory for 12 participants and 50 trials per condition in a within-subject design. In contrast, in a between-subjects design, 8 participants with 50 trials per condition are not enough to reach satisfactory power values. We conjecture that doubling the numbers of participants and trials will result in satisfactory power values; however, based on our data sets, we are not able to provide the required simulations. For the P3, we recommend a data set with at least 12 participants and 70 trials per condition, if one expects latency differences around 30 ms in a within-subject design. For smaller effect sizes, of course, more participants and/or more trials are required. To detect N2pc latency shifts of approximately 30 ms, 12 participants with 100 trials (50 for left and 50 for right stimulation) for each condition result in reasonable power for within-subject comparisons. For between-subjects comparisons, 16 participants with 100 trials are not sufficient to reliably detect latency differences of 30 ms. Here we recommend at least 24 participants with at least 100 trials per condition. Finally, power to detect latency shifts of almost 50 ms for the frequency-related P3 was not satisfactory for 18 participants with 35 trials in the infrequent and 95 trials in the frequent condition. In this case, increasing the number of trials may be problematic for practical reasons (e.g., increasing the number of infrequent trials to 70 would result in an increase to 190 frequent trials); thus researchers most likely will have to increase power by increasing the number of participants. Here we recommend at least 24 participants with no fewer trials than we applied to detect 50-ms latency shifts, but, of course, more participants are required the smaller the expected effect sizes.

Despite the fact that jackknifing had previously been shown to work well for the measurement of LRP onset latency, the present study was carried out because the important differences between LRPs and the other components examined here made it unclear whether jackknifing is an effective general-purpose technique for use with a variety of ERP components. Indeed, our simulations reveal that the jackknife approach is not always advantageous when combined with any technique. Because there is no formal theoretical way in which to determine for which cases the jackknife approach will estimate mean differences more or less correctly than the single-subject approach, we relied on simulation results. Based on these, we do consider the jackknife approach when combined with the

[3] We thank an anonymous reviewer for pointing to this possibility.

peak latency technique as inferior, because this combination produced rather low power for some data sets. Thus, if in some instance the peak latency technique seems especially appropriate, the jackknife approach should be avoided. However, due to the results with the new components, we think it is time to suggest that the superiority of jackknifing may be the general rule rather than the exception, when jackknifing is combined with either the relative criterion technique or with the fractional area technique. So we suggest that jackknifing should be considered as an appropriate analysis approach for new components as well.

In summary, the jackknife-based approach combined with the relative criterion technique or with the fractional area technique was never worse than the other methods, and, in general, these methods provided the most accurate estimates and the greatest statistical power, with no inflation of Type I error rate. This was true for simulations that evaluated onset latency for early sensory-perceptual components (auditory and visual N1) and later cognitive components (e.g., P3) that differed in amplitude, duration, and shape (e.g., with more or less defined onsets and peaks). It was also true for components that are isolated by computing difference waves from different electrode sites within trials (N2pc) and from the same electrode site across trial types (frequency-related P3). Finally, it was true for within-subject and between-subjects designs. Without doubt, further simulations analogous to those carried out here would strengthen confidence in the appropriateness of jackknifing combined with the relative criterion technique or with the fractional area technique with some new component. Even without such simulations, however, it seems reasonable to consider these methods to be the most appropriate analysis tool in any new, untested, situation, given the repeated superiority of these methods that has now been demonstrated with a wide variety of components in both within-subject and between-subjects designs.

REFERENCES

Alexander, J. E., & Polich, J. (1995). P300 differences between sinistrals and dextrals. *Cognitive Brain Research*, *2*, 277–282.

Arnell, K. M., Helion, A. M., Hurdelbrink, J. A., & Pasieka, B. (2004). Dissociating sources of dual-task interference using electrophysiology. *Psychonomic Bulletin & Review*, *11*, 77–83.

Brisson, B., & Jolicœur, P. (2007a). Cross-modal multitasking processing deficits prior to the central bottleneck revealed by event-related potentials. *Neuropsychologia*, *45*, 3038–3053.

Brisson, B., & Jolicœur, P. (2007b). Electrophysiological evidence of central interference on the control of visual–spatial attention. *Psychonomic Bulletin & Review*, *14*, 126–132.

Brisson, B., & Jolicœur, P. (2007c). A psychological refractory period in access to visual short-term memory and the deployment of visual–spatial attention: Multitasking processing deficits revealed by event-related potentials. *Psychophysiology*, *44*, 323–333.

Callaway, E., Halliday, R., Naylor, H., & Schechter, G. (1985). Effects of oral scopolamine on human stimulus evaluation. *Psychopharmacology*, *85*, 133–138.

Coles, M. G. H., Smid, H. G. O. M., Scheffers, M. K., & Otten, L. J. (1995). Mental chronometry and the study of human information processing. In M. D. Rugg & M. G. H. Coles (Eds.), *Electrophysiology of mind: Event-related brain potentials and cognition* (pp. 86–131). Oxford, UK: Oxford University Press.

Covington, J. W., & Polich, J. (1996). P300, stimulus intensity and modality. *Electroencephalography and Clinical Neurophysiology*, *100*, 579–584.

Curran, T., Hills, A., Patterson, M. B., & Strauss, M. E. (2001). Effects of aging on visuospatial attention: An ERP study. *Neuropsychologia*, *39*, 288–301.

De Beaumont, L., Brisson, B., Lassonde, M., & Jolicœur, P. (2007). Long-term electrophysiological changes in athletes with a history of multiple concussions. *Brain Injury*, *21*, 631–644.

Dell'Acqua, R., Jolicœur, P., Vespignani, F., & Toffanin, P. (2005). Central processing overlap modulates P3 latency. *Experimental Brain Research*, *165*, 54–68.

Dell'Acqua, R., Sessa, P., Jolicœur, P., & Robitaille, N. (2006). Spatial attention freezes during the attentional blink. *Psychophysiology*, *43*, 394–400.

Dimitrijevic, A., & Stapells, D. R. (2006). Human electrophysiological examination of buildup of the precedence effect. *NeuroReport*, *17*, 1133–1137.

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

Donchin, E., & Coles, M. G. H. (1988). Is the P300 component a manifestation of context updating? *Behavioral Brain Sciences*, *11*, 357–374.

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

Duncan-Johnson, C. C., & Donchin, E. (1977). On quantifying surprise: The variation of event-related potentials with subjective probability. *Psychophysiology*, *14*, 456–467.

Efron, B. (1981). Nonparametric estimates of standard error: The jackknife, the bootstrap, and other methods. *Biometrika*, *68*, 589–599.

Eimer, M. (1995). Event-related potential correlates of transient attention shifts to color and location. *Biological Psychology*, *41*, 167–182.

Eimer, M. (1996). The N2pc component as an indicator of attentional selectivity. *Electroencephalography and Clinical Neurophysiology*, *99*, 225–234.

Eimer, M., & Mazza, V. (2005). Electrophysiological correlates of change detection. *Psychophysiology*, *42*, 328–342.

Girelli, M., & Luck, S. J. (1997). Are the same attentional mechanisms used to detect visual search targets defined by color, orientation, and motion? *Journal of Cognitive Neuroscience*, *9*, 238–253.

Hansen, J. C., & Hillyard, S. A. (1980). Endogenous brain potentials associated with selective auditory attention. *Electroencephalography and Clinical Neurophysiology*, *49*, 277–290.

Hillyard, S. A., Hink, R. F., Schwent, V. L., & Picton, T. W. (1973). Electrical signs of selective attention in the human brain. *Science*, *182*, 177–179.

Jackson, P. R. (1986). Robust methods in statistics. In A. D. Lovie (Ed.), *New developments in statistics for psychology and the social sciences* (pp. 22–43). New York: British Psychological Society and Methuen.

Jemel, B., Schuller, A.-M., Cheref-Khan, Y., Goffaux, V., Crommelinck, M., & Bruyer, R. (2003). Stepwise emergence of the face-sensitive N170 event-related potential component. *NeuroReport*, *14*, 2035–2039.

Jentzsch, I., Leuthold, H., & Ulrich, R. (2007). Decomposing sources of response slowing in the PRP paradigm. *Journal of Experimental Psychology: Human Perception and Performance*, *33*, 610–626.

Johnson, R. Jr. (1986). A triarchic model of P300 amplitude. *Psychophysiology*, *23*, 367–384.

Jolicœur, P., Sessa, P., Dell'Acqua, R., & Robitaille, N. (2006a). Attentional control and capture in the attentional blink paradigm: Evidence from human electrophysiology. *European Journal of Cognitive Psychology*, *18*, 560–578.

Jolicœur, P., Sessa, P., Dell'Acqua, R., & Robitaille, N. (2006b). On the control of visual spatial attention: Evidence from human electrophysiology. *Psychological Research*, *70*, 414–424.

Kutas, M., McCarthy, G., & Donchin, E. (1977). Augmenting mental chronometry: The P300 as a measure of stimulus evaluation time. *Science*, *197*, 792–795.

Leuthold, H., & Sommer, W. (1998). Postperceptual effects and P300 latency. *Psychophysiology*, *35*, 34–46.

Luck, S. J. (1998). Sources of dual-task interference: Evidence from human electrophysiology. *Psychological Science*, *9*, 223–227.

Luck, S. J. (2005). *An introduction to the event-related potential technique*. Cambridge, MA: MIT Press.

Luck, S. J., Girelli, M., McDermott, M. T., & Ford, M. A. (1997). Bridging the gap between monkey neurophysiology and human perception: An ambiguity resolution theory of visual selective attention. *Cognitive Psychology*, *33*, 64–87.

Luck, S. J., & Hillyard, S. A. (1994). Spatial filtering during visual search: Evidence from human electrophysiology. *Journal of Experimental Psychology: Human Perception & Performance*, *20*, 1000–1014.

Magliero, A., Bashore, T. R., Coles, M. G. H., & Donchin, E. (1984). On the dependence of P300 latency on stimulus evaluation processes. *Psychophysiology*, *21*, 171–186.

Mangun, G. R. (1995). Neural mechanisms of visual selective attention. *Psychophysiology*, *32*, 4–18.

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

Miller, J., Patterson, T., & Ulrich, R. (1998). Jackknife-based method for measuring LRP onset latency differences. *Psychophysiology*, *35*, 99–115.

Miller, R. G. (1974). The jackknife—A review. *Biometrika*, *61*, 1–15.

Mordkoff, J. T., & Gianaros, P. J. (2000). Detecting the onset of the lateralized readiness potential: A comparison of available methods and procedures. *Psychophysiology*, *37*, 347–360.

Mosteller, F., & Tukey, J. (1977). *Data analysis and regression*. Reading, MA: Addison-Wesley.

Mulder, G. (1984). Stage analysis of the reaction process using brain-evoked potentials and reaction time. *Psychological Research*, *46*, 15–32.

Osman, A., Bashore, T. R., Coles, M., Donchin, E., & Meyer, D. (1992). On the transmission of partial information: Inferences from movement-related brain potentials. *Journal of Experimental Psychology: Human Perception and Performance*, *18*, 217–232.

Sable, J. J., Low, K. A., Maclin, E. L., Fabiani, M., & Gratton, G. (2004). Latent inhibition mediates N1 attenuation to repeating sounds. *Psychophysiology*, *41*, 636–642.

Smulders, F. T. Y., Kok, A., Kenemans, J. L., & Bashore, T. R. (1995). The temporal selectivity of additive factor effects on the reaction process revealed in ERP component latencies. *Acta Psychologica*, *90*, 97–109.

Tachibana, H., Aragane, K., Miyata, Y., & Sugita, M. (1997). Electrophysiological analysis of cognitive slowing in Parkinson's disease. *Journal of the Neurological Sciences*, *149*, 47–56.

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

Verleger, R. (1988). Event-related potentials and cognition: A critique of the context updating hypothesis and an alternative interpretation of the P3. *Behavioral Brain Science*, *11*, 343–427.

Verleger, R. (1997). On the utility of P3 latency as an index of mental chronometry. *Psychophysiology*, *34*, 131–156.

Verleger, R., Neukater, W., Kompf, D., & Vieregge, P. (1991). On the reasons for the delay of P3 latency in healthy elderly subjects. *Electroencephalography and Clinical Neurophysiology*, *79*, 488–502.

Vogel, E. K., & Luck, S. J. (2000). The visual N1 component as an index of a discrimination process. *Psychophysiology*, *37*, 190–203.

Vogel, E. K., & Luck, S. J. (2002). Delayed working memory consolidation during the attentional blink. *Psychonomic Bulletin & Review*, *9*, 739–743.

Vogel, E. K., Luck, S. J., & Shapiro, K. L. (1998). Electrophysiological evidence for a post-perceptual locus of suppression during the attentional blink. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 1654–1674.

Wascher, E. (2005). The timing of stimulus localization and the Simon effect: An ERP study. *Experimental Brain Research*, *163*, 430–439.

Woldorff, M. G., Gallen, C. C., Hampson, S. A., Hillyard, S. A., Pantev, C., Sobel, D., et al. (1993). Modulation of early sensory processing in human auditory cortex during auditory selective attention. *Proceedings of the National Academy of Sciences, USA*, *90*, 8722–8726.

Woodman, G. F., & Luck, S. J. (2003). Serial deployment of attention during visual search. *Journal of Experimental Psychology: Human Perception & Performance*, *29*, 121–138.

(Received March 12, 2007; Accepted August 22, 2007)