

# Reaction-time binning: A simple method for increasing the resolving power of ERP averages

RICCARDO POLI, CATERINA CINEL, LUCA CITI, AND FRANCISCO SEPULVEDA

Brain Computer Interfaces Laboratory, School of Computer Science and Electronic Engineering, University of Essex, Colchester, UK

## Abstract

Stimulus-locked, response-locked, and ERP-locked averaging are effective methods for reducing artifacts in ERP analysis. However, they suffer from a magnifying-glass effect: they increase the resolution of specific ERPs at the cost of blurring other ERPs. Here we propose an extremely simple technique—binning trials based on response times and then averaging—which can significantly alleviate the problems of other averaging methods. We have empirically evaluated the technique in an experiment where the task requires detecting a target in the presence of distractors. We have also studied the signal-to-noise ratio and the resolving power of averages with and without binning. Results indicate that the method produces clearer representations of ERPs than either stimulus-locked and response-locked averaging, revealing finer details of ERPs and helping in the evaluation of the amplitude and latency of ERP waves. The method is applicable to within-subject and between-subject averages.

**Descriptors:** ERP averaging, ERP signal-to-noise ratio, High-resolution averages, Reaction-time distributions, Variable-latency ERPs, Grand averages

While the study of *single-trial* Event Related Potentials (ERPs) has been considered of great importance since the early days of ERP analysis, in practice the presence of noise and artifacts has forced researchers to make use of *averaging* as part of their standard investigation methodology (Cobb & Dawson, 1960; Donchin & Lindsley, 1968; Handy, 2004; Luck, 2005).

Averaging is used in two ways in ERP analysis: to derive mean ERP waveforms for each subject taking part in an experiment and to compute averages of such waveforms (grand averages). There are essentially three classes of methods that are commonly used to resolve ERPs via averaging and a further class of methods where ERPs are reconstructed through the use of mathematical models. In this introductory section, we will start by reviewing these methods, discussing their strengths and weaknesses. We will then look at grand averaging and, finally, we will summarize the main ideas and contributions of this paper.

The authors would like to thank the Associate Editor (Dr Dean Salisbury) and the anonymous reviewers for their extremely useful comments and suggestions for improving this manuscript. This work was supported by the Engineering and Physical Sciences Research Council under [grant “Analogue Evolutionary Brain Computer Interfaces,” EP/F033818/1]; and the Experimental Psychological Society (UK) [grant “Binding Across the Senses”].

Address reprint requests to: Prof Riccardo Poli, Brain Computer Interfaces Laboratory, School of Computer Science and Electronic Engineering, University of Essex, Colchester, 304 ESQ, UK, E-mail: rpoli@essex.ac.uk

## Stimulus-locked Averaging

*Stimulus-locked averaging* requires extracting epochs from electroencephalogram (EEG) signals starting at stimulus presentation and averaging the corresponding ERPs. This is probably the oldest ERP analysis technique, dating back to the days of analogue averaging devices (Lindsley, 1968). Yet, it is still an effective means of investigation (e.g., see Kopp, Rist, & Mattler, 1996; Handy, 2004; Nieuwenhuis, Yeung, & Cohen, 2004).

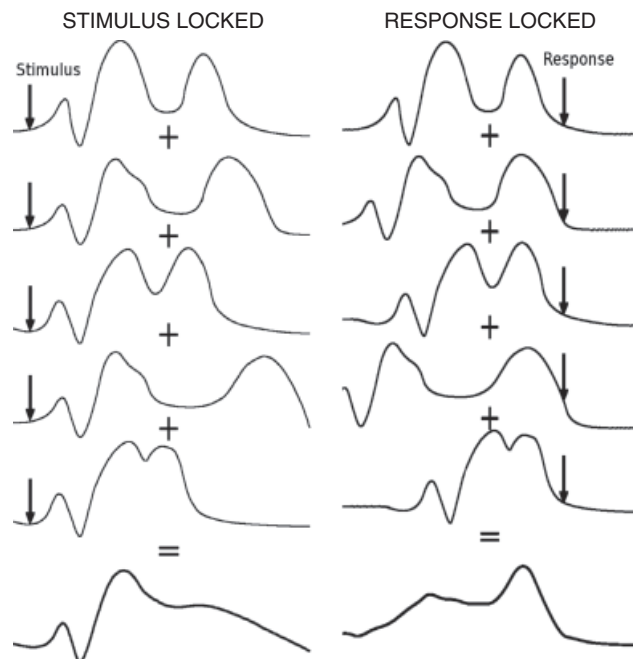
An important problem with this form of averaging is that any ERPs whose latency is not phase-locked with the presentation of the stimuli may be significantly distorted or may even completely disappear as a result of averaging (Spencer, 2004; Luck, 2005). This is because the average,  $a(t)$ , of randomly shifted versions of a waveform,  $w(t)$ , is the convolution between the original waveform and the latency distribution,  $\ell(t)$ , for that waveform, i.e.,  $a(t) = w(t) \star \ell(t) = \int w(t - \tau)\ell(\tau)d\tau$  (e.g., see Zhang, 1998). Given that latency distributions are non-negative and typically unimodal, this means that a stimulus-locked average can only show a smoothed (low-pass filtered) version of each variable-latency ERP. Furthermore, whenever the latency distribution of an ERP is unknown, the degree to which it will appear deformed in the average and in what ways it will be deformed are also unknown, hampering the interpretation of averages.

The problem is particularly severe when the task is relatively difficult, since the variability in the latency of endogenous ERPs and response times increase with the complexity of the task (Luck, 2005; Polich & Comerchero, 2003). In these cases, multiple endogenous variable-latency ERPs may appear as a single large smooth wave in the average; a synthetic example is shown in Figure 1 (left). This makes it difficult to infer true

brain area activity for any response occurring after the early exogenous potentials typically elicited by (and synchronized with) a stimulus.

### Response-locked Averaging

In experiments in which the task requires participants to provide a clearly identifiable response, *response-locked averaging* can be used as an alternative to stimulus-locked averaging to help resolve variable-latency ERPs that are synchronized with the response (e.g., see Luck & Hillyard, 1990; Keus, Jenks, & Schwarz, 2005; Spencer, 2004; Töllner, Gramann, Müller, Kiss, & Eimer, 2008). If participants are told to respond as soon as they have made a decision as to what action to take, then averaging should be expected to resolve waves which are at an approximately constant temporal position in relation to the response. The use of response-locked averages can be advantageous, for example, when effects of different experimental conditions on ERPs are caused by processes related to response selection, response preparation, or response inhibition, since these are likely to manifest themselves as response-locked ERPs (e.g., see Nieuwenhuis, Yeung, van den Wildenberg, & Ridderinkhof, 2003). In this case, however, the early responses associated and phase-locked with the stimulus will end up being blurred and hard to distinguish, since they are represented in the average by the convolution of their true waveform with the response-time distribution (Zhang, 1998). An example illustrating this problem is shown in Figure 1 (right).



**Figure 1.** Example of distortions produced by averaging: the five sample potentials at the top present two positive and one negative deflections each, which are phase-locked with a stimulus, as well as one positive ERP, which is of variable latency. Averaging them (plots at the bottom) preserves the exogenous ERPs when trials are stimulus-locked (left). This, however, turns the variable-latency ERPs into an inconspicuous plateau, which could easily be misinterpreted as a continuation of the preceding positive wave. A response-locked average (right), on the other hand, preserves the variable-latency endogenous ERP but smears out the details of early potentials turning them into a single, wide positive deflection.

Thus, in forced-choice experiments a researcher is presented with two alternative but often radically different or even conflicting representations of the same data: one based on stimulus-locked averaging and one based on response-locked averaging. Inferring whether a wave in the average represents a true effect or it is due to averaging biases can then be difficult. In addition, the deformations produced by blurring may lead to the incorrect evaluation of ERP parameters, such as the onset latency (which, in the average, reflects the fastest trials rather than typical ones). While one can qualitatively integrate the information provided by these two averages, and it is even possible to quantitatively morph them, it is unclear how reliable the result will be.

A key problem is that acquiring and *averaging more data does not help increase the fidelity of the reconstructed signals* because there is a *systematic error* (a distortion with a non-zero mean) in the averaging process. The lack of resolution for ERPs that are not phase-locked with external events is particularly problematic for difficult tasks that require several hundred milliseconds or even seconds to complete and that may involve multiple ERPs (e.g., related to stimulus evaluation, response selection, and action). This limits the applicability of stimulus-locked and response-locked ERP averaging for investigating processes taking place in the presence of richer and more realistic sets of stimuli and tasks.

### ERP-locked Averaging

A third alternative to resolve variable-latency ERPs is to attempt to identify them in each trial and estimate their latency. Then, shifting trials on the basis of estimated latencies and averaging may bring out the desired ERP from its noise background.

In some cases, simple techniques can be used to identify latencies of known waves. For example, P300s can be located by finding the largest positive deflection in a time window between 300 ms and, say, 800 ms after stimulus presentation (e.g., see Spencer, Abad, & Donchin, 2000) or the point at which the area under the signal in that time window has reached 50% of its maximum value (Luck, 2005).

An important issue about ERP-locked averaging is that most methods require *prior knowledge* about the ERP to be located. For example, one might need to tell an algorithm whether the ERP of interest is positive or negative, its approximate duration, and in what particular time window after stimulus presentation this is likely to occur. Without this information, automated detection algorithms have very little hope of finding the latency of the waves of interest. While such knowledge is often available, information can be contradictory. For example, the shape of ERPs may depend on whether one uses AC- or DC-coupled amplifiers and the degree to which pre-processing filters affect the frequency spectrum of such ERPs. Furthermore, the polarity and amplitude of an ERP may be reference- and electrode-dependent. Nonetheless, provided one is careful to refer to studies where experimental conditions closely match those of his or her own experiment, for relatively simple experimental conditions and ERPs, reasonably reliable knowledge to feed into an ERP latency-measuring algorithm can be found. However, how would we know what variable-latency ERPs will be present at different stages in the processing of a complex stimulus or the carrying out of a taxing cognitive task? Stimulus-locked or response-locked averaging might be unable to help in the identification of such ERPs. Also, if one hypothesizes the existence of a particular ERP and then runs a latency detection algorithm for it on a single-trial basis, it is very likely that the hypothesis will *appear* to be

corroborated irrespective of whether or not the ERP really exists. For example, if one locates positive peaks in random fragments of EEG and averages enough of them, the average will show a clear but entirely artefactual peak.

A related problem is that latency detection algorithms assume that the ERP of interest is present in every trial and we just need to find it. What if an ERP is not always elicited by the stimuli? The ERP might be, for example, dependent on whether a participant attended a stimulus, whether a participant was rested or tired, etc. (e.g., see Bonala, Boutros, & Jansen, 2008; Wagner, Rschke, Grzinger, & Mann, 2000). If an ERP was absent frequently, running a latency-measuring algorithm on trials where the ERP did not occur would inundate the averaging process with bias and noise.

An approach to average ERPs which accounts for trial-to-trial variability without requiring prior knowledge was introduced by Woody (1967). This is an adaptive filtering technique where the standard average is used as a starting template for an iterative process. For each epoch, the lag corresponding to the maximum covariance with this template is found. Then, each epoch is shifted by its lag and the shifted trials are averaged. The hypothesis is that this will result in a new template that is more accurate than the original one because ERPs were aligned. The process is iterated until a fixed-point is reached, i.e., no further shifts are required.

Wastell (1977) analyzed the accuracy of Woody's method, finding that the part of the signal that most closely matches the template may not be the ERP of interest. Recently, Thornton (2008) studied the behavior of Woody's filter as a function of the signal-to-noise ratio (SNR) in the data and found that below a SNR of 5 dB the method becomes unreliable. Since in real ERP recordings SNR values tend to be much worse than 5 dB, the filter may frequently lead to incorrect shifting of trials resulting in ERP-locked averages that significantly misrepresent reality. So, Woody's method gives accurate results only when the ERP of interest is large, sufficiently dissimilar from noise, and with latency distributions with relatively small standard deviations (Luck, 2005).

Naturally, it is possible to improve these techniques (e.g., see Thornton, 2008). However, all methods that realign trials based on ERP latencies are likely to suffer from a *clear-center/blurred-surround problem*. That is, after shifting trials based on an ERP's latencies, all instances of that ERP will be synchronized, effectively becoming fixed-latency elements. However, stimulus-locked ERPs will now become variable-latency ERPs. Also, all (other) ERPs that are phase-locked with some other event (e.g., the response), but not with the ERP of interest, will remain variable-latency. Not surprisingly, then, they will appear blurred and distorted in an ERP-locked average.

In summary, ERP-locked averaging is safe to use to reveal information about variable-latency ERPs that are known to exist, whose main characteristics have been identified using other methods and that are present in every trial being averaged. However, one needs to be very careful when using them as tools for identifying newly hypothesized ERPs.

### Model-based ERP Reconstruction

To overcome the problems of the methods reviewed above and better reconstruct ERPs, researchers have explored a variety of tools from statistics, signal processing, etc. All these methods make *strong assumptions on the definition and nature of the ERPs*

*to be reconstructed and on the nature of their interactions*. Below we review some key techniques.

Let us assume that the signal recorded in a forced-choice experiment is the sum of two ERPs—a stimulus-locked ERP,  $s(t)$ , and response-locked ERP,  $r(t)$ —and that the response-time does not affect their shape but only their relative position within a trial. Under these assumptions, it is possible to recover the two “true” ERPs from the response-locked average,  $a_r(t)$ , the stimulus-locked average,  $a_s(t)$ , and the response-time distribution,  $\rho(t)$  (Hansen, 1983; Zhang, 1998). The approach effectively involves jointly solving the two equations  $a_s(t) = s(t) + r(t) \star \rho(t)$  and  $a_r(t) = r(t) + s(t) \star \rho(-t)$  for  $s(t)$  and  $r(t)$  in the frequency domain and then anti-transforming the result. The technique has recently been extended (Yin, Zhang, Tian, & Yao, 2009) to deal with the case of experiments involving cues in addition to stimuli and responses. A potential problem for this technique is that it may be difficult to check the degree to which the assumptions it relies on are valid for a particular experiment. Also, the technique cannot recover variable-latency ERPs that are not phased-locked with an externally observable event: only partial information can be recovered and only under further strong assumptions.

Under a linear model of ERP interaction, when it is reasonable to assume that some ERPs are present with substantially the same amplitude and latency in two experiments while other ERPs are present only in one, it may be possible to isolate the former from the latter. For example, by adopting Kok's (1988) additive model of interaction between motor-related potentials (MRPs) and P300s, Salisbury, O'Donnell, McCarley, Nestor, Faux et al. (1994) were able to compute an average MRP waveform and to subtract MRP contamination from the P300's average waveform. The technique was later refined by Salisbury, Rutherford, Shenton, and McCarley (2001), who corrected the effects of MRPs on P300s on a trial-by-trial basis by carefully matching pairs of trials in two variants of an experiment according to the associated response time. An advantage of subtraction techniques of this type is that complex mathematical manipulations of the data are not required. Naturally, if the ERPs of interest are of variable latency, or variable-latency ERPs other than MRPs are present, the average of the recovered waveforms will still be affected by the low-pass filtering effects discussed above. Also, since the variance of the difference of stochastic variables is the sum of their variances, the process of subtracting ERPs (whether averaged or not) increases the noise affecting the data by a factor of  $\sqrt{2}$ , which may need to be compensated by the acquisition of more data.

Principal Component Analysis (PCA) has been suggested as a powerful statistical tool for the analysis of EEG and ERPs since the mid sixties (Streeter & Raviv, 1966; Donchin, 1966). PCA is based on the idea that the data are in fact a linear combination of “principal components” which need to be identified. PCA components are orthogonal, and they maximally account for the variance present in the data. Because of this, it is often possible to accurately represent the original data with a small set of components. Two forms of PCA are used in ERP analysis: one where one wants to find components that represent the covariance in the measurements taken at different electrodes, and one where one is interested in modelling the temporal variations in a signal. The latter—temporal PCA—is relevant in the context of this section.

Temporal PCA has been effective in identifying ERPs and in clarifying how ERPs vary as a result of changes in the stimuli, treatments, or subject groups (e.g., see Donchin, 1966; Do &

Kirk, 1999; Spencer et al., 2000; Dien, Spencer, & Donchin, 2003; Kayser & Tenke, 2006). However, the technique makes strong assumptions (see Donchin & Heffley, 1978). Firstly, PCA is linear: it assumes that ERPs do not interact. Secondly, it assumes that the major sources of variance (the principal components) are orthogonal; so different, but correlated ERPs may end up being represented by a single component. Thirdly, the technique implicitly assumes that only the amplitudes of ERPs vary, not their latency; when this is not the case, the PCA component associated to an ERP may totally misrepresent reality (Donchin & Heffley, 1978). Thus, the use of PCA in ERP analysis requires significant care, and there is evidence that results may be misleading (e.g., see Beauducel & Debener, 2003). Variable-latency ERPs cannot be properly resolved with this technique.

Independent Component Analysis (ICA) (e.g., see Hyvärinen, Karhunen, & Oja, 2001) has also seen considerable popularity in the analysis of EEG and ERPs (Makeig, Bell, Jung, & Sejnowski, 1996; Makeig, Jung, Bell, Ghahremani, & Sejnowski, 1997; Makeig, Westerfield, Jung, Covington, Townsend et al., 1999; Jung, Makeig, Westerfield, Townsend, Courchesne et al., 2001; Makeig, Westerfield, Jung, Enghoff, Townsend et al., 2002). If a set of signals is the result of linearly superimposing some statistically independent sources, ICA can decompose the signals into their primitive sources. These are called “independent components.” When ICA is applied to the signals recorded at different electrodes on the scalp, it can separate important sources of EEG and ERP variability. This can then be exploited, for example, for removing artifacts. The use of ICA for reconstructing ERPs with varying latency has also been trialled (Jung et al., 2001). In the presence of variable-latency ERPs, the method tends to allocate different ICA components to different ERPs if they originate from different areas. ICA-based reconstruction of variable-latency ERPs requires that the different ICA components that capture separate ERPs be appropriately temporally shifted so as to realign the components. Then, at least in principle, anti-transforming the shifted ICA components together with any non-shifted ones should reconstruct a signal where all ERPs are fully resolved. However, the ICA component alignment process is manual and to some extent arbitrary, as is the identification of the number of ERPs that need reconstructing. So, the method must be guided by prior knowledge and results may present significant inter- and intra-experimenter variability. Also, it is hard to interpret what exactly the final resulting “average” represents, since it is essentially a morph between stimulus-locked and ERP-locked averages. Finally, the application of ICA to ERP analysis is based on strong assumptions: the linearity of the brain as a conduction medium, the statistical independence of the sources of electrical activity, the fixed position of such sources, the absence of conduction delays in the brain, and the non-Gaussianity of the statistical distributions of the sources (Makeig et al., 1996, 1997, 1999; 2002; Jung et al., 2001). It may be difficult to verify to what extent these assumptions are tenable for a specific experiment.

#### Grand Averages and Averages Across Subjects

In ERP analysis, a grand average is simply an average of the average waveforms obtained on a subject by subject basis. The purpose of grand averages is the reduction of the noise that may affect single-subject averages and the identification of the commonalities between such averages. A defect of grand averages is that they may not represent the waveforms recorded from indi-

vidual subjects well (Luck, 2005). One reason for this is that, even if all subjects present the same sequence of ERPs in a given condition, such ERPs—particularly variable-latency ones—are likely to have different latencies in the averages of different subjects. Thus, averaging such averages will produce low-pass filtering effects similar to those affecting ordinary ERP averages.

While grand averages are the most widespread technique to combine evidence across subjects, it is important to also consider the alternative of simply averaging all trials pertaining to a certain condition irrespective of subject. This is because the two strategies address different questions: grand averaging answers the question of what the ERPs of a *typical subject* in a certain condition look like, averaging across subjects addresses the question of what the *typical waveform* for the ERPs recorded in a particular condition is.

Grand averages and averages across subjects are mathematically very similar. Let  $v_{ij}(t)$  be the  $j$ -th ERP recorded for subject  $i$ . The subject’s average is given by  $a_i(t) = \sum_{j=1}^{m_i} \frac{v_{ij}(t)}{m_i}$ , where  $m_i$  is the number of trials for that subject. Then, the grand average is given by

$$a_g(t) = \sum_{i=1}^n \frac{a_i(t)}{n} \quad (1)$$

where  $n$  is the number of subjects. An average across subjects, instead, is given by

$$a_a(t) = \sum_{i=1}^n \sum_{j=1}^{m_i} \frac{v_{ij}(t)}{N} = \sum_{i=1}^n \frac{m_i}{N} a_i(t), \quad (2)$$

where  $N = \sum_{i=1}^n m_i$  is the total number of trials. Clearly  $a_g(t)$  and  $a_a(t)$  can differ only if the number of trials for one or more subjects differs from the mean number of trials, i.e.,  $m_i \neq \frac{N}{n}$ . Even when this happens, though, significant differences are likely only in the presence of small samples and large individual differences in both  $m_i$  and  $v_{ij}(t)$  across subjects.

Both grand averages and averages across subjects can be computed for stimulus-locked, response-locked, and ERP-locked averages.

#### Contributions of this Paper

It is clear from the survey presented above that a more precise and direct way of identifying variable-latency ERPs as well as measuring their latency and amplitude is needed. This need is particularly pressing in the presence of complex and realistic tasks where precise knowledge may not be available about which ERPs are present and how their amplitudes and latencies are affected by particular conditions.

In this paper, we propose a simple technique which we believe can achieve this: *binning trials based on their recorded response time and then computing bin averages*. This has the potential of solving the problems of stimulus-locked, response-locked, and ERP-locked averages, effectively reconciling them. In particular, response-time binning can significantly improve the resolution with which variable-latency waves can be recovered via averaging. The reason is simple.

The idea is that if one selects out of a dataset all those epochs where a participant was presented with qualitatively identical stimuli and gave the same response within approximately the same amount of time, it is reasonable to assume that similar internal processes will have taken place (we will call this a cognitive homogeneity assumption). So, within those trials, ERPs that would normally have a widely variable latency might be

expected to, instead, present a much narrower latency distribution. Thus, if we bin epochs on the basis of stimuli, responses, and response times, we should find that, *for the epochs within a bin, the stimulus, the response, as well as fixed- and variable-latency ERPs are much more synchronized* than if one did not divide the dataset. Averaging such epochs should, therefore, allow the rejection of noise while at the same time reducing also the undesirable distortions and blurring associated with averaging and avoiding the complexities, strong assumptions, or manual labor involved in the application of model-based reconstruction methods. Response-time binning and averaging should result in clearer descriptions of brain activity *without the need for prior knowledge* of the phenomena taking place and ERPs elicited in response to the stimuli. In this paper, we describe our implementation, analysis, and evaluation of this technique.

Many studies on the relationship between reaction times and the amplitude and the latency of ERPs have been reported in the literature (see, for example, McCarthy & Donchin, 1981; Kutas, McCarthy, & Donchin, 1977; Donchin, Ritter, & McCallum, 1978). Typically they rely on the trial-by-trial measurement of the amplitude and/or latency of ERPs and the statistical analysis of their covariance with corresponding response times (see Childers, Perry, Fischler, Boaz, & Arroyo, 1987). In a smaller fraction of the studies, however, trials were divided up into broad groups by reaction time, e.g., fast vs slow responses (as in Makeig et al., 1999; Woodman & Luck, 1999), and were then averaged. We are also aware of one study (Roth, Ford, & Kopell, 1978) where ERP trials were grouped by response-time quartiles and one (Gratton, Coles, Sirevaag, Eriksen, & Donchin, 1988) where trials were grouped using four predefined response-time intervals. While this prior work presents some similarities with what we propose here, there are also significant methodological and philosophical differences. We discuss them below.

First, the subdivision of trials into groups based on response time in previous work is virtually always motivated by the desire to measure the amplitude or latency of one specific wave (e.g., the P300, as in Roth et al. (1978)) in each group and then to relate such measures to the corresponding response times. Here, instead, we are proposing the use of response-time binning not just to better understand the relationship between specific waves and reaction times, but chiefly as a method to identify and resolve such waves in the first place, thanks to the increased resolution it provides.

Second, as we will see, response-time binning increases the effective resolution of bin averages in an inverse proportion to the bin size. So, although for testing purposes here we used four bins, as in Roth et al., 1978 and Gratton et al., 1988 (we throw away our fourth bin being largely made up of outliers), we propose to use as many bins as the cardinality of the dataset can support.

Third, we suggest that to truly benefit from the resolution enhancements provided by response-time binning one needs to use it in forced-choice experiments *and* after trials have already been divided up by stimulus type and response, to ensure that the bins are as homogeneous as possible. Failing to do so may render the technique pointless from the resolving-power viewpoint. In our tests, we use a forced-choice set-up, and we do not simply divide the trials into ‘Correct’ and ‘Incorrect,’ but into ‘True Positives,’ ‘True Negatives,’ ‘False Positives,’ and ‘False Negatives.’ On the contrary, for instance, the experimental set-up in Roth and colleagues, 1978 was not one of forced-choice (the absence of a response within 800 ms from the stimulus was taken as a negative response). So, the trials that were averaged in the last quartile were not homogeneous.

Fourth, while binning is useful also in the case of short response times, we propose that it is really in experiments requiring the processing of complex stimuli or the performance of complex tasks with correspondingly longer reaction times that the binning technique can provide the biggest advantages over other ERP averaging methods. However, previous work involving the use of response-time bins has mainly focused on simple tasks. For example, in Roth et al. (1978), reaction times in the first and fourth quartiles were 366 ms and 540 ms, respectively, while in Gratton et al., 1988, four 50 ms-wide bins covered the range 150–349 ms.

Finally, we should note that, whenever one divides up a noisy dataset into subsets and then studies the subsets separately, each subset contains fewer trials and, thus, inference of true effects and ERPs from the noise is harder. So, there is a trade-off between the desire to gain more precise information by averaging signals that are more homogeneous and the loss of precision due to the reduced noise rejection associated to smaller sets. Here we analyze this trade-off via an evaluation of how dividing up trials by reaction-time affects SNR. To the best of our knowledge, an analysis of this kind has never been reported in the literature. Furthermore, for the first time we will formally relate the use of bins to the resolution with which fixed latency and variable-latency ERPs can be recovered via averaging by connecting response-time binning to the theory proposed in Hansen (1983) and Zhang (1998). So, the paper also fills significant theoretical gaps.

## Methodology

### *Response-time Binning*

In ERP experiments, EEG signals are partitioned into epochs. In our tests with response-time binning we used epochs starting at the onset of a stimulus and lasting 1200 ms.

Naturally, response-time binning requires deciding how many bins to use and how wide each bin should be. Given that binning reduces the number of epochs contributing to ERP statistics (e.g., bin averages) thereby increasing the noise affecting them, it may be best to start with a small number of bins to check if this reveals previously undetected regularities. If this is the case, and more resolution is desired, one can divide the dataset more finely. If noise becomes a problem, one can test participants for longer or over multiple sessions with the reasonable expectation that the additional data will further increase our knowledge of a phenomenon (which is not necessarily the case for standard averages when waves with varying latencies are present). Thus, to demonstrate our technique, we divided epochs into three main bins.

In many conditions, response times have highly skewed distributions with long upper tails. Therefore, unless one is specifically interested in studying waves corresponding to unusually long response times, to avoid their smearing effect on averages it is important to discard events in the extreme right tails of response-time distributions. In this paper, we chose to discard the trials falling in the rightmost 10% quantile (i.e., the 10th decile) of the distributions.

Once these anomalous data have been removed, we are faced with an important dilemma. In principle, it would seem desirable to create response-time bins equally temporally spaced, i.e., all of the same width. This would tend to give the same temporal resolution to bin averages. However, because response-time distributions are skewed, doing so would create bins with very unequal numbers of epochs in them, resulting in bin averages being

affected by radically different noise levels. A better alternative from this point of view is to consider bins which correspond to equal areas under the distribution. Because of the shape of response-time distributions, this may produce bins of unequal sizes and averages with different effective resolutions.

In our work, we adopted this approach. Thus, from the 90% of the trials left after removing the 10th decile of the distribution, we created three bins: one gathering the lower 30% of response times (*bin 1*), one for the middle 30% (*bin 2*), and one for the longer 30% (*bin 3*).

#### **Recordings, Artifact Removal, and Trimmed Averaging**

EEG signals were acquired using a BioSemi ActiveTwo system (BioSemi, Amsterdam, The Netherlands) with 64 pre-amplified DC-coupled electrodes spaced evenly over the scalp. Additional electrodes were placed at the earlobes for off-line referencing, at the left and right external canthi to record horizontal electro-oculogram (HEOG), and infra-orbitally to record vertical electro-oculogram (VEOG). Signals were acquired at 2,048 samples per second, were then bandpass-filtered between 0.15 and 40 Hz and, finally, were down-sampled to 512 samples per second.

Effects of eye blinks and vertical components of saccades were reduced by using the time-domain linear regression between each channel and the VEOG. That is, we subtracted from each EEG channel a proportion of the signals recorded by the two VEOG channels; the proportion was obtained by computing the correlation between the EEG signals recorded at each electrode with the VEOG signals and dividing by the VEOG's power (Verleger, Gasser, & Mcks, 1982; Luck, 2005).

We then applied to each bin an artifact rejection procedure which involved computing the first ( $Q_1(t)$ ) and third ( $Q_3(t)$ ) quartiles of the voltages at each time step across all the epochs in a bin. The procedure then removed all epochs where the signal was outside the range

$$[Q_1(t) - 1.5(Q_3(t) - Q_1(t)), Q_3(t) + 1.5(Q_3(t) - Q_1(t))]$$

for more than 10% of the samples in an epoch. The remaining epochs in each bin were then used to compute statistics, e.g., for averaging. This procedure was also used to remove artifacts before computing statistics when data were not binned.

To further reduce the effect of outliers, instead of simply averaging trials, unless otherwise stated, we used *40%-trimmed averages*. These are robust measures of central tendency that are less sensitive to outliers than the ordinary mean (Huber, 1981). They have been shown to provide significant increases in reliability compared to ordinary averages in ERP and event-related desynchronization analysis (Gasser, Mcks, & Khler, 1986; Burgess and Gruzelier, 1999; Rousselet, Husk, Bennett, & Sekuler, 2008). Trimmed averages are computed as follows. For each time step, the voltages recorded in the epochs in a bin are sorted, and the upper and lower 40% are discarded. The remaining central (and somehow most representative) 20% of the voltages are then averaged. The process is repeated for each time step.

#### **Experiment's Methodology**

To evaluate response-time binning, we modified a forced-choice experiment designed by Esterman, Prinzmetal, and Robertson (2004) where the task requires detecting a target colored letter in the presence of distractor letters of different colors. Participants, stimuli, and procedure are described below.

*Participants.* Six students from the University of Essex (average age: 24 years; five females; one participant was left-

handed). All had normal or corrected-to-normal vision and had normal color vision.

*Stimuli.* On each trial, participants were presented with a four-letter string. The first and last letters were always 'S.' Of the two middle letters, one was always an 'O,' while the other was either an 'L' or an 'X.' Letters subtended an angle of  $1.19^\circ$  vertically. The horizontal gap between letters also subtended  $1.19^\circ$ . The first and last letters were always white, while the color of the two middle letters could either be red, green, or blue, but never the same color. The background was black.

Each letter string was randomly presented in one of four regions of the display. These extended from the center of the screen to its top-left, top-right, bottom-left, and bottom-right corners, respectively. The horizontal displacement of the inner edge of a string with respect to the centre of the screen varied between  $5.27^\circ$  and  $10.47^\circ$ . The vertical displacement of the string was always identical to the horizontal displacement.

*Procedure.* In the experiment, participants had to decide whether or not, on each display, a target letter was presented. The target was always an 'L' of a specific color.

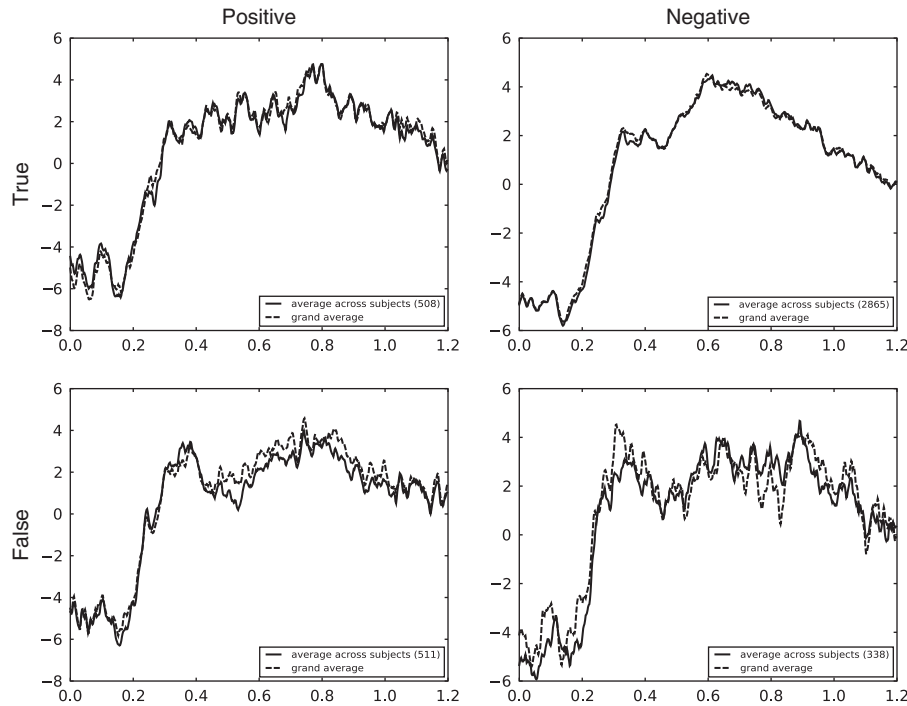
The experiment was divided into blocks of 40 trials each. The target was present in 20% of the trials. At the beginning of the experiment, participants were told the color of the target letter. Every six blocks, the target color was changed so as to test each participant on each target color on an equal number of trials. Target color order was counterbalanced across subjects.

To control the level of difficulty of the experiment, the stimuli to be presented in each block were carefully chosen in relation to the frequency of targets and non-targets as well as the possibility of being deceived by letters having one (but not both) the features of the target. The color and letter combinations used for the two middle letters in the stimulus string in the 40 trials in each block were as follows: eight trials with the target (an 'L' of a specific color) and an 'O' of a non-target color; eight trials with an 'O' of the target color and an 'L' of a non-target color; eight trials where an 'O' and an 'L' both of a non-target color were presented; four trials where an 'X' with the target color and an 'O' with a non-target color were presented; four trials where an 'X' had the non-target color and an 'O' had the target color; eight trials where an 'X' and an 'O' were presented, both in a non-target color. In every block, trial order was randomized.

A white dot was always visible at the center of the display. At the beginning of each trial, the dot was replaced by a fixation cross for 500 ms, and then the letter string briefly appeared in one of the quadrants. Participants were instructed to gaze at the white dot/cross and to try not to move their eyes when the stimulus string was presented. The string was displayed for a duration which was adjusted at the end of each block of the experiment, according to the percentages of correct responses in the block. The objective was to keep a subject's accuracy between 75% and 90%. This procedure ensured stimulus presentation was fast enough to make target detection relatively difficult, while at the same time discouraging participants from guessing too often.

The duration of the stimulus display varied between 50 ms and 150 ms (with intermediate steps that were multiples of the inverse of our computer screen refresh rate, which was 60 Hz). All participants started at 150 ms. The most frequent presentation times were 83 and 100 ms.

The horizontal and vertical displacements of the letter string were also changed in relation to performance. The first block's



**Figure 2.** Comparison between grand averages and averages across subjects for true positive (top left), false positive (bottom left), true negative (top right), and false negative (bottom right) for ERPs recorded on channel Cz.

displacements were  $5.27^\circ$ . If a participant's accuracy was too high, displacements were increased in following blocks to  $8.01^\circ$ , and then, if necessary, to  $10.47^\circ$ .

Participants gave their responses by pressing the left button of a mouse with the index finger for 'Yes' responses and the right button with the middle finger for 'No' responses. Each response was followed by an interval of 1 sec, after which the next trial started.

After a practice session, each participant completed six blocks with each target color for a total of 18 blocks.

## Results

In this section, we will empirically evaluate the binning technique using the data collected in the experiment described above. Trials were divided into four categories (true positives, true negatives, false positives, and false negatives) according to whether the target was present or absent and whether the response was 'Yes' or 'No.' Unless otherwise stated, the results for each category are based on cumulating the trials of all subjects. So, most of the ERP averages we show are across subjects. In our experiment, these are qualitatively very similar to grand averages as illustrated in Figure 2 for our four conditions. We will also report some single-subject results to illustrate the applicability of the method to the study of within-subject ERP variability.

### Response Time Statistics and Bins

We show the response-time distributions recorded in our experiments for these four conditions in Figure 3 (note that amplitudes have been normalized so that the curves are proper density functions, i.e., the area under each curve is unitary; abscissas are in seconds). For each condition we created three bins, each containing 30% of the distribution (the right-most 10% of the dis-

tribution was discarded). Bin boundaries are shown as vertical lines in Figure 3.

The medians and standard deviations, estimated using the standard robust estimator provided by  $1.4826 \times$  the median absolute deviation from the median (or MAD for short), for the whole distribution as well as the bins in each condition are shown graphically in Figure 3 and numerically in Table 1. Table 1 also reports the ranges of response times associated to each bin in different conditions. The number of trials in each class are shown in Table 2.

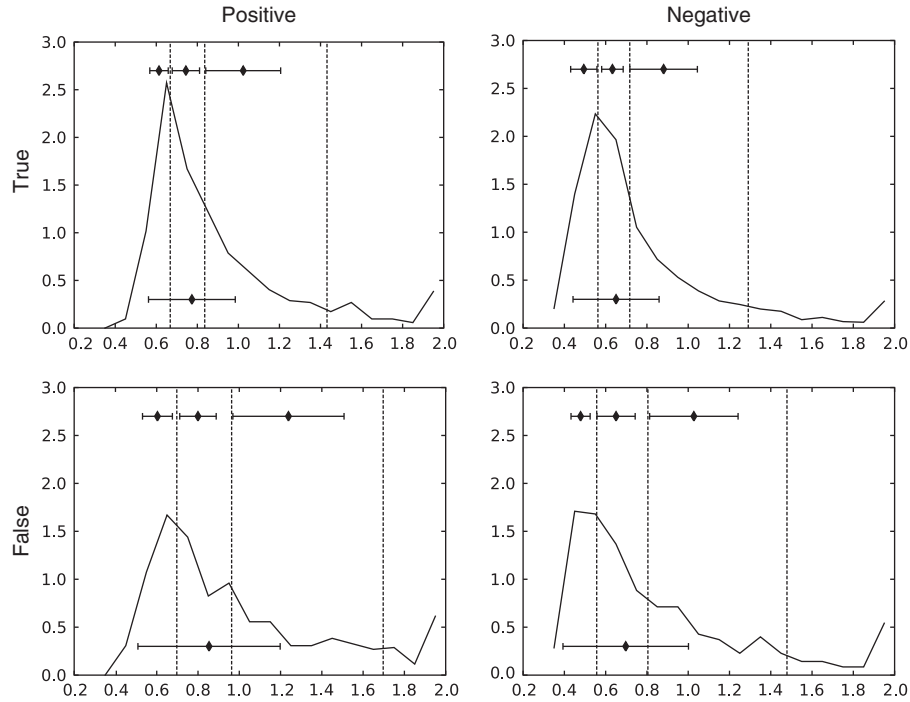
Although the distributions in Figure 3 have some similarities, they are in fact pairwise statistically different because of the large size of our samples. This is clearly shown by the results of the Kolmogorov-Smirnov test for distributions presented in Table 3. This suggests that dividing the data into four categories by response and presence/absence of target makes sense.

### Stimulus-locked and Response-locked Averaging

We used 40%-trimmed averages to represent the typical behavior of ERPs for different stimulus-response pairs (i.e., true/false positives/negatives) both for the bins and for the whole sample.

Let us start by looking at what happens if we compute stimulus-locked and response-locked averages of the four classes without binning. Figure 4 shows the results of this process for the electrodes Cz and Pz, for our four conditions. In the plots on the left, the stimulus onset is at 0 ms. For easier comparison, the response-locked-average plots on the right were independently shifted so that the stimulus is also at 0 ms. Let us analyze these plots.

We can clearly see from Figure 4 that, as expected, the stimulus-locked averages on the left show the early ERPs clearly. Conversely, there is very little detail on patterns of activity after the first 400 ms, i.e., preceding and immediately following the response. The situation is completely reversed when we consider



**Figure 3.** Response time distributions for true positive (top left), false positive (bottom left), true negative (top right), and false negative (bottom right) trials. Response times of 2 s or longer have been grouped in the rightmost point of each distribution. The vertical lines within each plot represent the boundaries of the bins produced by our binning method. In each plot, medians and standard deviations (estimated via the corrected median-absolute deviation) are also shown both for the bins (upper part of the plot) and for the overall distribution (lower part of the plot).

the response-locked averages on the right. Here the early potentials are effectively impossible to discern, while waves of activity in the proximity of the response and their individual differences appear to be well resolved.

The differences between the stimulus-locked and the response-locked representations of brain activity are generally very large (we will see an illustration of this in the next section). So, it may be difficult to integrate the two averages into a unified interpretation. In general, it is hard to discern how far away from a synchronizing event (whether stimulus or response) we can go before the waves shown in an average are just artifacts.

**Table 1.** Median, Standard Deviation, and Ranges of Reaction Times for Bins and Whole Dataset in Different Conditions

Medians and Standard Deviations				
	True Positives	True Negatives	False Negatives	False Positives
All	0.77 ± 0.21	0.65 ± 0.21	0.70 ± 0.30	0.85 ± 0.34
Bin 1	0.61 ± 0.05	0.49 ± 0.06	0.48 ± 0.05	0.60 ± 0.07
Bin 2	0.74 ± 0.07	0.63 ± 0.05	0.65 ± 0.09	0.80 ± 0.09
Bin 3	1.02 ± 0.18	0.88 ± 0.16	1.03 ± 0.21	1.24 ± 0.27

Response-time Ranges				
	True Positives	True Negatives	False Negatives	False Positives
All	0.00–2.00	0.00–2.00	0.00–2.00	0.00–2.00
Bin 1	0.00–0.67	0.00–0.56	0.00–0.56	0.00–0.70
Bin 2	0.67–0.84	0.56–0.72	0.56–0.81	0.70–0.96
Bin 3	0.84–1.43	0.72–1.29	0.81–1.48	0.96–1.70

*Note:* Values are in seconds. Standard deviations are estimated via corrected median-absolute deviations.

### Bin-based Averaging

Let us now look at the averages obtained using response-time binning for the different conditions.

Let us start from the true negative trials. The first row of Figure 5 shows the stimulus-locked averages for channels Cz and Pz. It is immediately apparent how much crisper than in Figure 4 (left) the different ERPs are when using bins. Also, it is clear how different response times are in fact associated with different amplitudes and latencies in ERPs, particularly for the late ERPs following the exogenous responses. We should note, however, that bin 3 has a much wider response-time distribution than the other two bins (see Figure 3). The relative lack of late activity in bin 3 can, therefore, be partly attributed to residual latency jitter.

The improvement in the effective resolution is also confirmed by the averages for the false negatives shown in Figure 5 (second row). Using binning, we can now see how qualitatively different the ERPs produced in the presence of an incorrect response can be. For example, it is clear how bin 3 deviates from the others: in Cz we can see an early potential which is not present in the averages for the other bins, while the large positive wave occurring in those bins between approximately 500 ms and 700 ms is totally absent from the third bin. Why is this happening? We can easily hypothesize explanations, but we will not attempt to

**Table 2.** Number of Trials in Each Class, in the Bins, and in the Whole Sample (Before Artifact Rejection)

	True Positives	True Negatives	False Negatives	False Positives
All	521	2967	521	351
Bins	156	890	156	105

**Table 3.** Comparison of Response Time Distributions of Different Classes of Trials with Kolmogorov-Smirnov Test

	True Positives	True Negatives	False Negatives	False Positives
True Positives	N/A	0	$3.83 \times 10^{-13}$	0.00028
True Negatives	0	N/A	0.0013	0
False Negatives	$3.83 \times 10^{-13}$	0.0013	N/A	$5.27 \times 10^{-10}$
False Positives	0.00028	0	$5.27 \times 10^{-10}$	N/A

Note: The table reports the  $p$  value corresponding to each pair of distributions.

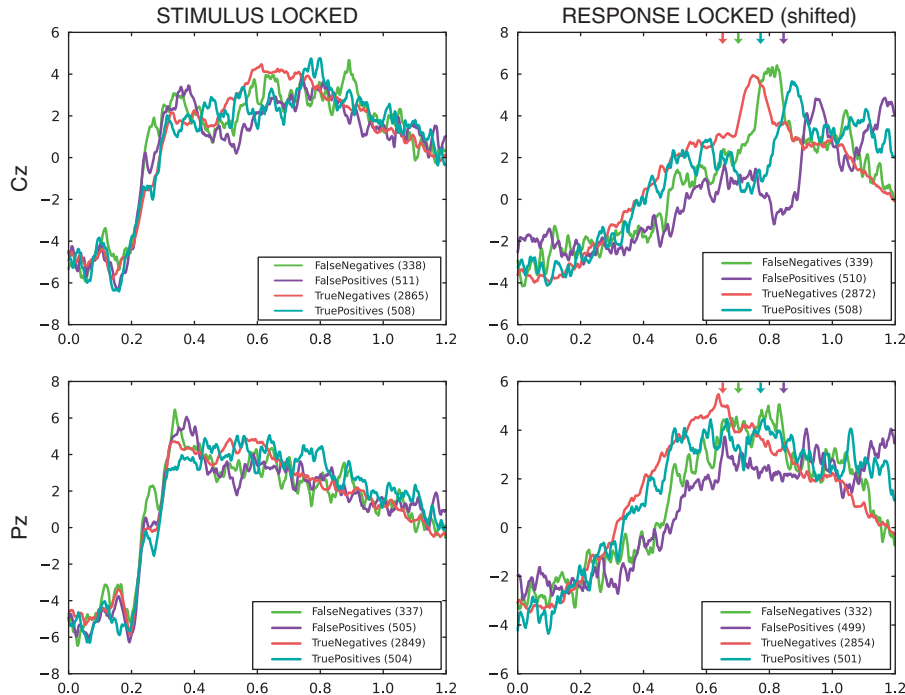
interpret these findings in this work. We should stress, however, that it is really the binning technique that has made it possible to ask these questions in the first place, by identifying otherwise undetectable differences.

As shown in Figure 5, the false positives present some similarities with the true negatives. These include, for example, the effective absence of a positive wave between approximately 500 ms and 700 ms in the average for bin 3.

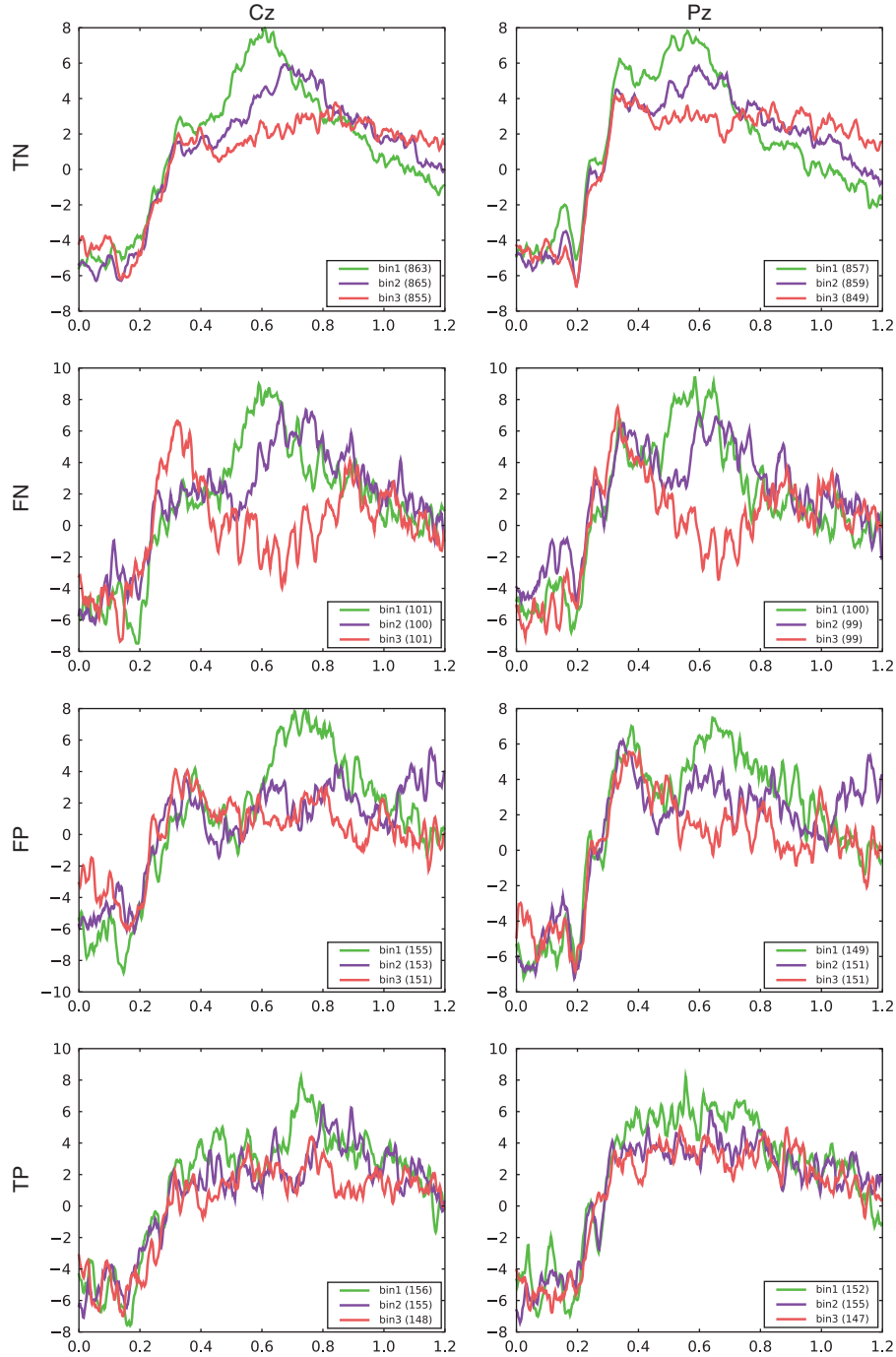
If we look at the averages for the true positives (bottom of Figure 5), we find that bin-average differences are much reduced, with only bin 1 showing an amplitude elevation between 300 ms and 800 ms. This suggests that the events taking place in the correct recognition of a target vary much less with the reaction time than for other categories. So, in a sense, the binning technique is telling us that in this case ordinary stimulus-locked averages can be trusted more than in the other cases. This is a general application of the binning technique. Whenever the plots for the averages of different bins coincide in a particular range of times, we should expect ordinary averages to be reasonably reliable for that range of times and vice versa.

In principle, one can average the trials in a bin either aligning them on the stimulus or on the response. As we mentioned earlier, stimulus-locked and response-locked averages can accurately resolve only waves which are phase locked with the corresponding reference event and it is, therefore, difficult to integrate the information they provide into a unified picture. As an illustration of the large differences between the two averages, in Figure 6 (top) we have superimposed the stimulus-locked and response-locked averages recorded in Cz for the True Negatives. So, it seems reasonable to ask whether binning can reduce the discrepancies in the information provided by the two averaging techniques.

Unsurprisingly, when bins are narrow, aligning the epochs in a bin based on stimulus onset or response produces very similar averages, as illustrated in Figure 6 (rows 2 to 4 left) for the true negatives for channel Cz. Note how similar the response-locked and stimulus-locked averages are for bins 1 and 2. This is common for all conditions. Only in bin 3 can we see discrepancies between the two averages. The reason is that, despite our removing the 10% of the distribution corresponding to the longest response times, bin 3 still has a much bigger response-time variance than the other two bins. Averaging biases will, therefore, manifest themselves also in bin 3, albeit to a lesser degree than in the absence of binning. It is then not surprising to see that for that bin the early ERPs are only well captured by the stimulus locked average, and vice versa. Note, however, that by showing differences in the two plots, the binning technique reveals that if one wants to study more precisely what happens in unusually long trials, the response time distribution needs to be divided more finely. Of course, since this reduces the number of trials present in each bin, one needs to test enough subjects and each subject for long enough (e.g., in multiple sessions) to ensure noise levels are sufficiently low.



**Figure 4.** Comparison of stimulus-locked (left) and response-locked (right) ERP averages recorded in our experiments for True Positive, True Negative, False Positive, and False Negative trials for electrode sites Cz and Pz. Times are in seconds. The arrows in the plots on the right indicate the median response time for each condition.



**Figure 5.** Trimmed stimulus-locked averages for True Negative (TN), False Negative (FN), False Positive (FP), and True Positive (TP) trials for channels Cz (left) and Pz (right).

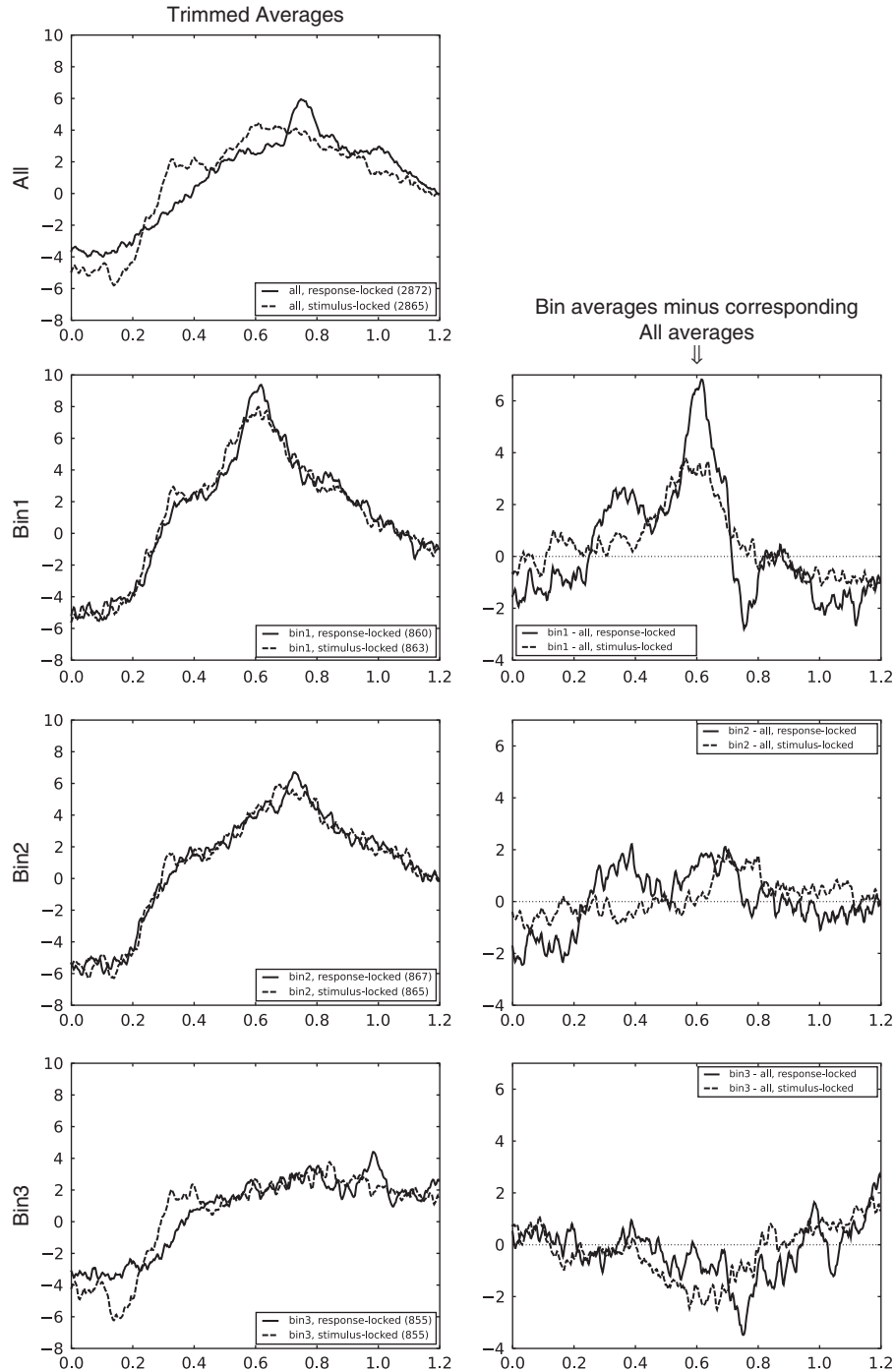
Overall, we can see that binning effectively brings the two main ways of studying ERP—stimulus-locked and response-locked averaging—closer, effectively unifying them for narrow bins.

On the right of Figure 6, we show plots of the difference between bin averages and the corresponding average obtained using all trials. For all bins, absolute differences of  $2 \mu\text{V}$  or more are present over periods of several hundred of milliseconds, particularly in the central region of the epochs. In that region, relative errors of 30% or more are common across all bins, with bin 1 showing differences of over  $6 \mu\text{V}$ , which correspond to relative

errors of nearly 70%. This suggests that a large proportion of the variance in ERP averages is actually accounted for by latency jitter.

#### **Reliability of Bin Averages**

The results presented in the previous section are of a qualitative nature. The question of the degree to which averages constructed via response-time binning are effective at resolving and properly representing ERPs needs to be addressed more formally. Let us start by checking whether observed differences are statistically significant. This can be done as follows.

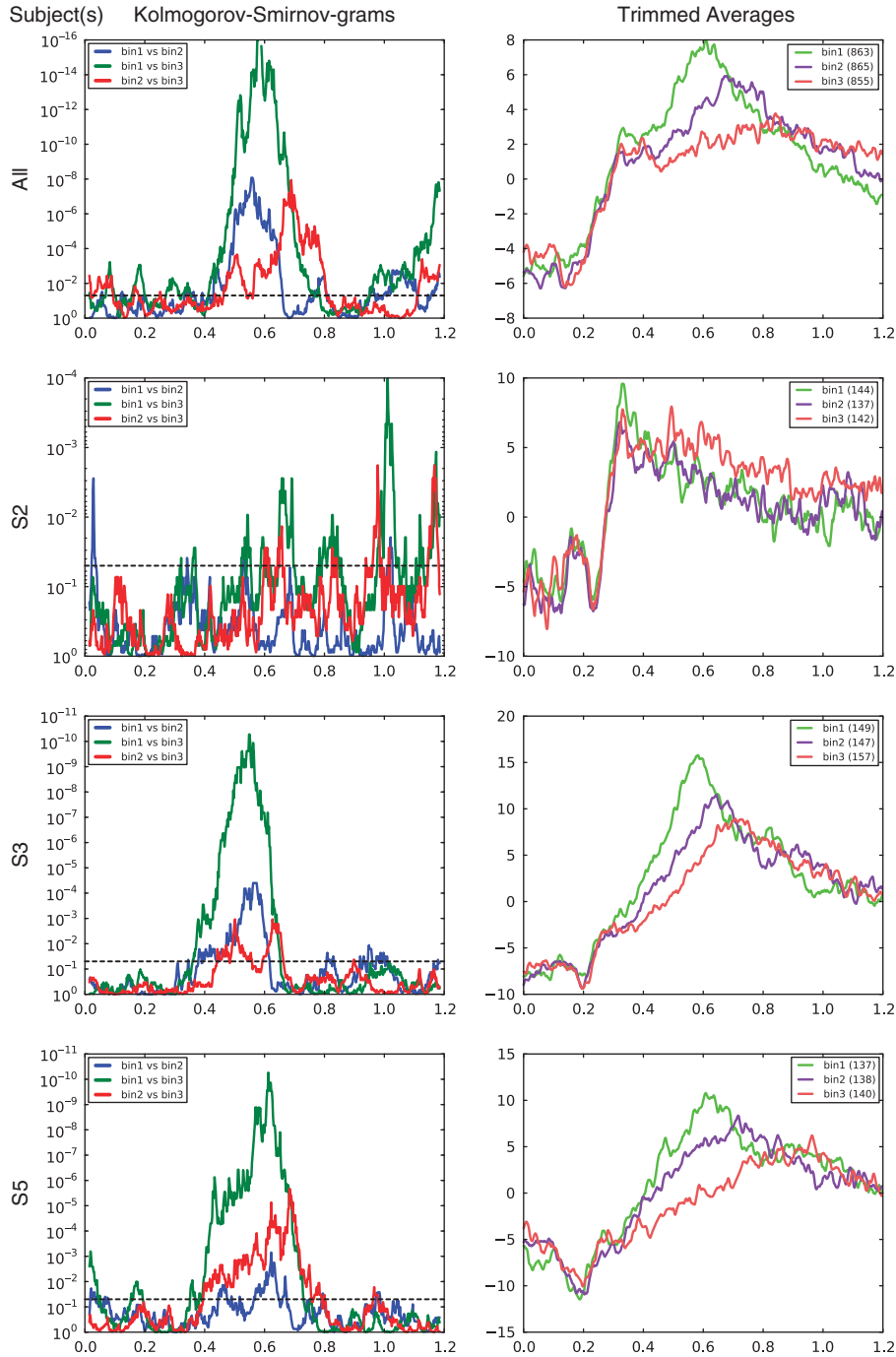


**Figure 6.** Comparison between stimulus-locked and response-locked 40% trimmed averages recorded in Cz for the True Negatives (our largest class) in the absence (top) and in the presence of response time binning (rows 2 to 4 left). The plots on the right show the difference between bin averages and the corresponding average obtained using all trials.

If we focus on one particular time step, we can treat each bin as a univariate sample of the amplitudes recorded at that particular time step in the epochs in the bin. We can then use the Kolmogorov-Smirnov test to check whether the samples in pairs of bins might be drawn from the same distribution. Since single-trial amplitudes are very noisy, ERP amplitudes are rarely estimated by looking at a single sample. So, instead of passing to the test the amplitudes of a specific sample, we can use amplitude averages taken over small intervals centered around

the time of interest in each trial. The  $p$  values obtained via the test when comparing bin amplitudes at a specific time will then reveal whether differences at that time are significant.

For example, if we look at the false positives and channel Cz across all subjects and compare bins 1 and 2, bins 1 and 3, and bins 2 and 3, we find that amplitude differences are all highly statistically significant between 550 ms and 600 ms ( $p = 5.43 \times 10^{-7}$ ,  $1.11 \times 10^{-16}$  and  $1.51 \times 10^{-3}$ , respectively), but are not in the interval 850 ms—900 ms ( $p = .53$ ,  $.50$ , and  $.23$ , respectively).



**Figure 7.** Analysis of statistical significance of amplitude distribution differences: (left)  $p$  values recorded in the Kolmogorov-Smirnov-grams (see text) for ERP amplitude differences observed in different bins for channel Cz and for the true negative class (note the inversion of the vertical scale where higher corresponds to smaller  $p$  values) and (right) bin trimmed averages (provided for reference) for all subjects together (top row) and for three typical subjects analyzed independently. The horizontal dashed lines in the KS-gram plots represent the 5% significance level.

A comprehensive representation of the intervals where amplitude differences between bins are significant is provided by what we could call a *Kolmogorov-Smirnov-gram* (or KS-gram for short), i.e., a plot of the  $p$  values obtained when sliding a time window over the trials and running the Kolmogorov-Smirnov test on the average amplitudes recorded in the window in a pair of bins. A diagram showing the KS-grams obtained for our three bins for channel Cz and for the true negative trials across all subjects is shown in Figure 7 (top left). The 5% significance level is represented by the horizontal dashed line in the figure. The KS-

grams in the figure were computed using a 30 ms-wide sliding window. For reference, Figure 7 (top right) shows the averages for the bins.

As one can easily see, all plots in Figure 7 (top left) are on the ‘statistically significant’ side of the diagram for a large proportion of an epoch. More precisely, the KS-gram for bins 1 and 3 shows that there are statistically significant amplitude differences between the two bins for 36.3% of the epoch; the KS-gram for bins 1 and 2 indicates significant differences for 59.0% of the epoch; and ERP amplitudes in bins 2 and 3 are statistically sig-

nificantly different for 57.5% of the epoch. This cannot be explained by chance: binning by response-time must be capturing significant regularities in the ERPs evoked during the experiment. Furthermore, the major deviations in the bin averages observable in Figure 7 (top right) are all highly statistically significant. Additionally, bin-to-bin comparisons show that the early potentials which one would not expect to be heavily modulated by condition and response-time are indeed mostly below the significance threshold.

The results we reported above were based on cumulating the trials of all subjects. However, the binning technique can be applied also on a subject-by-subject basis. The plots in rows 2 to 4 on the left of Figure 7 show KS-grams for three typical subjects (S2, S3, and S5). The corresponding plots on the right show their bin averages. The KS-grams and bin averages obtained in the case of S3 and S5 (as well as S4 and S6, not reported) are qualitatively very similar to those obtained across all subjects. That is, we see that ERP amplitude distributions are significantly different across bins for a large proportion of the epochs, despite the fact that single-subject bins contain only 1/6 of the total dataset. In the case of S2 (and S1, not reported), instead, we find that differences are much smaller and only occasionally statistically significant. Upon inspection of the error rates for these subjects, we found that they had adjusted to the low target frequency in the experiment (20%) and tended to respond ‘No’ significantly more often than average. So, it is likely that they used a different strategy than the other subjects, which may explain the presence of the large positive ERP phased-locked with the stimulus presentation and immediately following the exogenous ERPs. These subjects were also characterized by the effective absence of variable-latency ERPs. This in turn led to the lack of significant differences between the bins as highlighted by their KS-grams.

An important question concerns the effects that dividing up a data set into bins based on response time has on noise. In the section entitled *Contributions of this Paper*, we suggested that there is a trade-off between the desire to gain precise information by averaging more homogeneous signals and the loss of precision due to the reduced noise rejection associated to smaller sets. Here we want to check this hypothesis.

To address this issue, we measured the Signal-to-Noise Ratio of the averages for different sets of trials using the technique developed by Schimmel (1967). That is, for each dataset and at each time step, we averaged the even-index and odd-index epochs in the dataset separately, obtaining the signals  $av_e(t)$  and  $av_o(t)$ , respectively. We then estimated the SNR of the mean as follows:

$$\text{SNR} = \frac{\text{RMS}\left(\frac{av_e(t)+av_o(t)}{2}\right)}{\text{RMS}\left(\frac{av_e(t)-av_o(t)}{2}\right)}$$

where RMS (for Root Mean Square) of a set of values is defined as the square root of the mean of the squares of the values in the set. The SNR values obtained for different stimulus-response groups (across all subjects) are shown in the first column of Table 4. The corresponding values for the bins are reported in columns two to four.

Let us consider these results in detail. Firstly, we should note that the SNR for the false negatives is substantially better than for the other classes. This is not surprising since this is by far our largest category, as we showed in Table 2. Also, it is not surprising to see that as we move from the ‘All’ column to the ‘Bin’ columns we see a drop in SNR. However, what is really important

**Table 4.** Signal-to-Noise Ratio for the Different Sets of Trials Evaluated According to the Schimmel (1967) Method (See Text) for Channel Pz

	All	Bin 1	Bin 2	Bin 3	Average SNR drop by class
True Positives	18.4 dB	16.8 dB	14.3 dB	11.8 dB	4.1 dB
True Negatives	22.5 dB	22.4 dB	19.7 dB	20.1 dB	1.8 dB
False Negatives	15.6 dB	12.8 dB	12.9 dB	12.5 dB	2.9 dB
False Positives	16.5 dB	15.0 dB	11.7 dB	11.5 dB	3.8 dB

to verify is whether the drop is any different from what we would expect if we randomly split a dataset into subsets.

Because the bins include 30% of the total sample, the theoretical SNR drop we would expect to see in the presence of random bins is  $20\log_{10}(\sqrt{0.3})$ , i.e.,  $-5.2$  dB, as we move from the first column to the second, third, and fourth in Table 4. However, the average drop in SNR when going from the whole group to the bins varies from 1.8 dB to 4.1 dB, with an average of 3.1 dB. The smallest SNR loss is associated with bin 1, where on average SNR drops by only about 1.5 dB. All this is possible because some of the variance present in the averages in the absence of binning is in fact due to variable-latency ERPs that effectively act as noise on the mean. With binning, instead, there is much less variability associated with variable-latency waves.

So, not only the reduction in the biases of averaging brought about by response-time binning results in an improvement in resolution, but it is also responsible for the SNR on the mean remaining significantly higher than expected despite the large sample-size reduction due to binning.

### Analysis of Resolution

In this section we will relate response-time binning to the work by Hansen (1983); Zhang (1998), thereby formally clarifying the reasons why this averaging technique increases the resolution with which ERPs can be recovered.

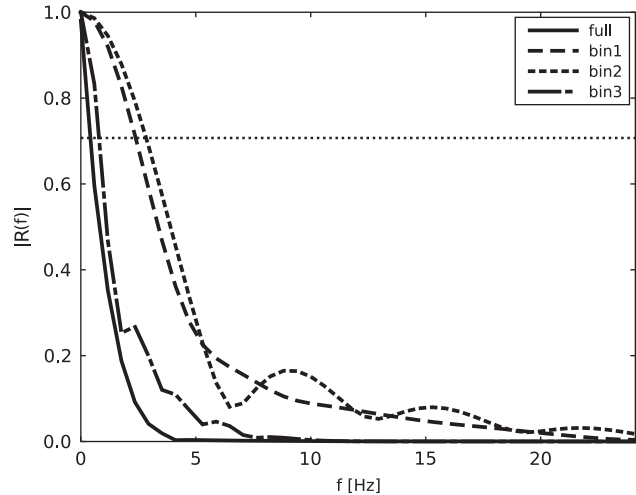
#### Resolution of Stimulus-locked and Response-locked ERPs

At first, let us make the same assumptions as in (Hansen, 1983; Zhang, 1998). Let us assume that there are two additive ERPs in the signals recorded in a forced-choice experiment—a stimulus-locked ERP,  $s(t)$ , and a response-locked ERP,  $r(t)$ —and that  $\rho(t)$  is the response-time density function. Under the further assumption that response times do not affect the shape of these ERPs, the stimulus-locked average can be expressed as  $a_s(t) = s(t) + r(t) \star \rho(t)$ , where  $\star$  is the convolution operation. A similar equation can be written for the response-locked average  $a_r(t)$ .

Let us consider in what ways binning by response time would affect this result. Let us define a function  $\delta(x)$  that returns 1 if  $x$  is true, and 0 otherwise. Then  $\delta(r_1 \leq t < r_2)$  can be seen as a membership function for the trials belonging to a bin characterized by response times within the interval  $[r_1, r_2)$ . Thus, the product  $\delta(r_1 \leq t < r_2)\rho(t)$  represents the distribution of response times within the bin. This can be turned into a probability density function by dividing it by  $\int_{r_1}^{r_2} \rho(t) dt$ .

It is then clear that the stimulus-locked bin average, which we denote as  $a_s^{[r_1, r_2)}(t)$ , is given by

$$a_s^{[r_1, r_2)}(t) = s(t) + r(t) \star \rho^{[r_1, r_2)}(t)$$



**Figure 8.** Amplitude of the frequency response of the response time distribution for all the true negatives recorded in our experiment and corresponding frequency responses for the three bins. The horizontal line in the figure represents the standard 3 dB attenuation line. The intersection between this line and each curve represents the cut-off frequency of the corresponding low-pass filter.

where

$$\rho^{[r_1, r_2]}(t) = \frac{\delta(r_1 \leq t < r_2)\rho(t)}{\int_{r_1}^{r_2} \rho(t) dt}$$

is the convolution kernel responsible for  $r(t)$  appearing blurred in the average. So, in order to understand whether  $a_s^{[r_1, r_2]}(t)$  provides a better representation of  $r(t)$  than  $a_s(t)$ , we need to analyze the differences between the distributions  $\rho^{[r_1, r_2]}(t)$  and  $\rho(t)$ .

Apart from a scaling factor, the key difference between the two is that  $\rho^{[r_1, r_2]}(t)$  is the product of  $\rho(t)$  and a rectangular windowing function,  $\delta(r_1 \leq t < r_2)$ . In the frequency domain, therefore, the spectrum of  $\rho^{[r_1, r_2]}(t)$ , which we denote with  $\mathcal{R}^{[r_1, r_2]}(f)$ , is the convolution between the spectrum of  $\rho(t)$ , denoted as  $\mathcal{R}(f)$ , and the spectrum of a translated rectangle,  $\Delta(f)$ . This is a scaled and rotated (in the complex plane) version of the *sync* function (i.e., it behaves like  $\frac{\sin(f)}{f}$ ). The function  $|\Delta(f)|$  has a large central lobe whose width is inversely proportional to the bin width,  $r_2 - r_1$ . Thus, when convolved with  $\mathcal{R}(f)$ ,  $\Delta(f)$  behaves as a low-pass filter. Therefore,  $\mathcal{R}^{[r_1, r_2]}(f) = \mathcal{R}(f) \star \Delta(f)$  is a smoothed and enlarged version of  $\mathcal{R}(f)$ . In other words, while  $\rho^{[r_1, r_2]}(t)$  is still a low-pass filter, it has a higher cut-off frequency than  $\rho(t)$ .

We illustrate this effect in Figure 8 for the class of true negatives. The figure shows the amplitude of the frequency response,  $|\mathcal{R}(f)|$ , of the response time distribution as well as the frequency responses for bins 1, 2, and 3,  $|\mathcal{R}^{[r_1, r_2]}(f)|$ . These were computed via the discrete Fourier transform of the distributions  $\rho(t)$  and  $\rho^{[r_1, r_2]}(t)$  obtained from the raw data. To improve the accuracy of the result, we derived high resolution representations of these distributions by using the Parzen window method (Parzen, 1962).

We should note how all the frequency responses shown in Figure 8 are characteristic of low-pass filters. The narrowest of all is the one associated with the full response time distribution (i.e., when no binning is performed). This has a cut-off frequency of 0.42 Hz. This implies that without binning a stimulus-locked

average can reproduce without significant distortion only response-locked ERPs which represent extremely slow potentials occurring over a period of perhaps 1 sec or more.

The second lowest cut-off frequency, 0.79 Hz, is associated with bin 3. It is not surprising to see that bin 3 has still a very low cut-off frequency, because this bin is the widest of all three, covering a large portion of the long upper tail of the response time distribution. Despite this, however, binning still has effectively *doubled the resolution of averaging for this bin*.

The representation of  $r(t)$  in bin averages is even less deformed than the representation of the same ERP in the ordinary average for bins 1 and 2, which have cut-off frequencies of 2.38 Hz and 2.84 Hz, respectively. For these bins, therefore, stimulus-locked averaging has *improved the resolving power on response-locked ERPs by six or seven times*. Thus, bin averages can reliably resolve features of response-locked ERPs down to durations of the order of 100–200 ms.

The narrower the bin, the smaller the deformations. So, if a higher resolution is required, one just needs to use narrower bins and to acquire correspondingly more trials. In fact, *for sufficiently small bins, the bin average is an unbiased estimator of the true ERP*. To illustrate this, let us imagine that we pick a bin which is so narrow that we can consider  $\rho(t)$  constant on it. So,  $\rho^{[r_1, r_2]}(t) = \delta(r_1 \leq t < r_2)/(r_2 - r_1)$ . Then, if we take the limit for the bin size  $(r_2 - r_1) \rightarrow 0$ , we get that  $\rho^{[r_1, r_2]}(t)$  approaches more and more a Dirak delta function. So, from the properties of the convolution operator we obtain:

$$\lim_{(r_2 - r_1) \rightarrow 0} a_s^{[r_1, r_2]}(t) = s(t) + r(t).$$

All of the properties discussed in this section also hold for binned response-locked averages.

### Resolution of Variable-latency ERPs

Let us consider a more general case. Let us assume that there are three additive ERPs in the signal recorded in a forced-choice experiment: the  $s(t)$  and  $r(t)$  waves mentioned above, and a variable-latency ERP,  $v(t)$ . Let  $R$  be a stochastic variable representing the response time in a trial;  $\rho(t)$  is its density function. Similarly, let  $L$  be a stochastic variable representing the latency of the ERP  $v(t)$  and let  $\ell(t)$  be the corresponding density function (or latency distribution). As above, let us further assume that response time and latency do not affect the shape of these ERPs. Under these assumptions, we obtain the following equation for the stimulus-locked average  $a_s(t)$ :

$$a_s(t) = s(t) + v(t) \star \ell(t) + r(t) \star \rho(t). \quad (3)$$

A similar equation can be written for the response-locked average  $a_r(t)$ .

Zhang (1998) considered a special version of this equation and showed that if *stimulus-locked and response-locked ERPs,  $s(t)$  and  $r(t)$ , are absent* and if the latency,  $L$ , and the lag,  $(R - L)$ , between the variable-latency ERP  $v(t)$  and the response are *statistically independent*, then some information about  $v(t)$  can be recovered from the knowledge of  $a_s(t)$ ,  $a_r(t)$ , and  $\rho(t)$ . In particular, one can find the amplitude of the Fourier transform of  $v(t)$ . Because the phase information cannot be recovered, however, the reconstruction of  $v(t)$  is not possible (see Zhang, 1998, for more details).

An interesting question is whether different assumptions on the relation between  $R$  and  $L$  might in fact allow one to derive more information about  $v(t)$ , while, perhaps, being more psychophysically tenable. Our objective in this section,

however, is more modest. Starting from Equation (3), we want to see how binning affects the resolution with which  $v(t)$  is represented in a stimulus-locked average.

Let us start by considering the most general conditions possible. Let  $L$  and  $R$  be described by an unspecified joint density function  $p(l, r)$ . So, the latency and response-time distributions are marginals of this joint distribution, i.e.,

$$\ell(l) = p(l, r) dr \quad \text{and} \quad \rho(r) = p(l, r) dl.$$

Note that by the definition of conditional density function, we also have that

$$p(l, r) = p(r|l)\ell(l) \quad \text{and} \quad p(l, r) = p(l|r)\rho(r)$$

where  $p(r|l)$  is the pdf of  $R$  when  $L = l$  and  $p(l|r)$  is the pdf of  $L$  when  $R = r$ .

Focusing our attention on the subset of the trials falling within the response-time bin  $[r_1, r_2]$ , i.e., such that  $r_1 \leq R < r_2$ , changes the joint distribution of  $L$  and  $R$  into

$$\begin{aligned} p^{[r_1, r_2]}(l, r) &= \frac{\delta(r_1 \leq r < r_2) \times p(l, r)}{\int_{r_1}^{r_2} p(l, r) dr dl} \\ &= \frac{\delta(r_1 \leq r < r_2) \times p(l, r)}{\int_{r_1}^{r_2} \int p(l, r) dl dr} \\ &= \frac{\delta(r_1 \leq r < r_2) \times p(l, r)}{\int_{r_1}^{r_2} \rho(r) dr}, \end{aligned}$$

where the  $\delta$  function zeroes the distribution outside the strip  $[r_1, r_2]$  and the denominator normalizes the result so that  $p^{[r_1, r_2]}(l, r)$  integrates to 1.

The marginal of this distribution with respect to  $l$  gives us the response time distribution for the response-time bin  $[r_1, r_2]$ :

$$\begin{aligned} \rho^{[r_1, r_2]}(r) &= \int p^{[r_1, r_2]}(l, r) dl = \frac{\delta(r_1 \leq r < r_2) \int p(l, r) dl}{\int_{r_1}^{r_2} \rho(r) dr} \\ &= \frac{\delta(r_1 \leq r < r_2) \rho(r)}{\int_{r_1}^{r_2} \rho(r) dr}, \end{aligned}$$

which confirms the definition we provided previously. More interesting is the marginal of the distribution  $p^{[r_1, r_2]}(l, r)$  with respect to  $r$ , which gives us the latency distribution for the trials in the response-time bin  $[r_1, r_2]$ :

$$\begin{aligned} \ell^{[r_1, r_2]}(l) &= \int p^{[r_1, r_2]}(l, r) dr = \frac{\int_{r_1}^{r_2} p(l, r) dr}{\int_{r_1}^{r_2} \rho(r) dr} = \frac{\left( \int_{r_1}^{r_2} p(r|l) dr \right) \ell(l)}{\int_{r_1}^{r_2} \rho(r) dr} \\ &= \frac{Pr\{r_1 \leq R < r_2 | l\} \ell(l)}{\int_{r_1}^{r_2} \rho(r) dr}. \end{aligned}$$

These two marginals are important because, analogously to what we did in the previous section, we can express the stimulus-locked bin average as follows:

$$a_s^{[r_1, r_2]}(t) = s(t) + v(t) \star \ell^{[r_1, r_2]}(t) + r(t) \star \rho^{[r_1, r_2]}(t).$$

So, these marginals determine in what ways and to what extent  $v(t)$  and  $r(t)$  appear deformed and blurred in the average. Because we have already analyzed the effects of the convolution with  $p^{[r_1, r_2]}(t)$  on the resolution of  $r(t)$ , in the rest of this section we will concentrate on analyzing  $\ell^{[r_1, r_2]}(t)$ .

The key difference between  $\ell(l)$  and  $\ell^{[r_1, r_2]}(l)$ , apart from a scaling factor, is that  $\ell^{[r_1, r_2]}(l)$  is the product of  $\ell(l)$  and a windowing function,  $w^{[r_1, r_2]}(l) = Pr\{r_1 \leq R < r_2 | l\}$ . In the frequency domain, therefore, the spectrum of  $\ell^{[r_1, r_2]}(l)$ , which

we denote with  $\mathcal{L}^{[r_1, r_2]}(f)$ , is the convolution between the spectrum of  $\ell(l)$ , denoted as  $\mathcal{L}(f)$ , and the spectrum of the window,  $w^{[r_1, r_2]}(f)$ . All we know about the function  $w^{[r_1, r_2]}(l)$  is that it can never be negative, being in fact a probability. Therefore, its spectrum  $w^{[r_1, r_2]}(f)$  must have a non-zero component at  $f=0$ . However, this does not necessarily imply that  $w^{[r_1, r_2]}(f)$  behaves as a low-pass filter. So, in general we cannot say for sure whether  $\mathcal{L}^{[r_1, r_2]}(f)$  is wider than  $\mathcal{L}(f)$ , which would imply that binning increases the resolution of  $v(t)$  in the average. As we will see below, however, under mild assumptions on the relationship between  $R$  and  $L$ , this is actually the case. Let us consider two cases.

*Case of deterministic functional dependency between  $R$  and  $L$ .* In the *Contribution* section, we put forward the following cognitive homogeneity assumption: if one considers those epochs where a participant was presented with qualitatively similar stimuli and gave the same response within approximately the same amount of time, it is reasonable to assume that similar internal processes will have taken place. Under this assumption, fixed- and variable-latency ERPs will appear much more synchronized than if one looked at an undivided dataset. The cognitive homogeneity assumption effectively implies that, within a stimulus/response class when  $R$  takes a particular value, the value of  $L$  is also approximately determined and vice versa.

To ease our mathematical analysis, let us idealize this assumption imagining that  $R = g(L)$  where  $g$  is some unknown deterministic function. Note that this assumption is the exact opposite of the independence assumption of Zhang (1998) since in our model ( $R$ - $L$ ) is dependent on  $L$  via the relation  $(R - L) = g(L) - L$ .

Because of the dependency between  $R$  and  $L$ , we have that  $Pr\{r_1 \leq R < r_2 | l\}$  is 1 if  $g(l) \in [r_1, r_2]$  and 0 otherwise. So,

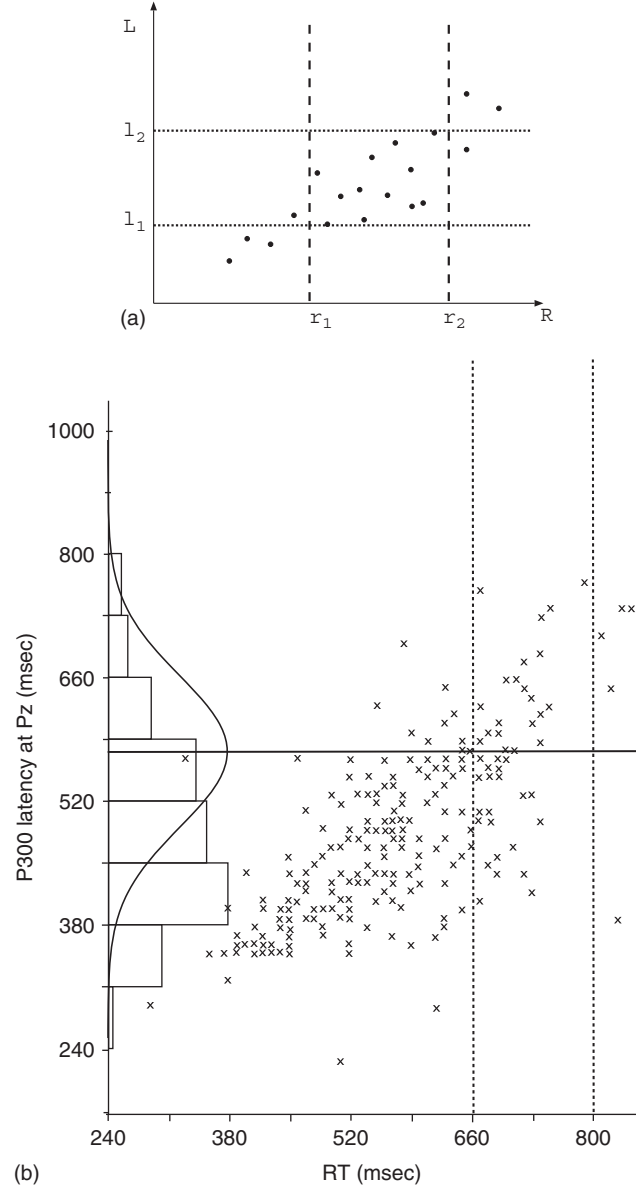
$$\ell^{[r_1, r_2]}(l) = \frac{\delta(g(l) \in [r_1, r_2]) \ell(l)}{\int_{r_1}^{r_2} \rho(r) dr}.$$

If additionally  $g$  is monotonic (which implies invertibility), this simplifies to

$$\ell^{[r_1, r_2]}(l) = \frac{\delta(\min(g^{-1}(r_1), g^{-1}(r_2)) \leq l < \max(g^{-1}(r_1), g^{-1}(r_2))) \times \ell(l)}{\int_{r_1}^{r_2} \rho(r) dr}.$$

Thus, in the bin average, the variable-latency ERP  $v(t)$  appears distorted via the convolution with a version of the latency distribution  $\ell(t)$  that has been passed through a rectangular window. As discussed earlier, rectangular windowing in the time domain corresponds to convolving the spectrum of  $\ell(t)$  with a sinc function. This has the effect of broadening the spectrum. Therefore, the convolution with  $\ell^{[r_1, r_2]}(t)$  blurs  $v(t)$  less than the convolution with  $\ell(t)$  does. Thus, under the assumptions we made above, *the resolution of variable-latency ERPs is enhanced in bin averages.*

*Case of stochastic dependency between  $R$  and  $L$ .* Binning increases the resolution of response-locked ERPs because it considers a narrower range of response times. In the model studied in the previous section, this benefit was also available for variable-latency ERPs since discarding trials whose response time was outside a particular range corresponded to rejecting variable-latency ERPs whose latency is outside some interval. So, it is reasonable to expect that something like this would still happen even if  $L$  wasn't a deterministic function of  $R$ , as long as there was a sufficiently strong correlation between  $R$  and  $L$ . Indeed, in such a case, in a scatterplot of the  $(R, L)$  pairs



**Figure 9.** (a) Imaginary scatterplot of response times ( $R$ ) vs latencies ( $L$ ). (b) Response times vs latency scatterplot for P300s reconstructed from Kutas et al., 1977 (p. 794, Figure 2, left) for the ‘accurate condition.’

associated to all the trials in a dataset, we would find that the data cloud tends to align (to a degree that depends on how strong the correlation between  $L$  and  $R$  is) along a line, such as the straight line obtained via linear regression. Picking a subset of the trials corresponding to  $R$  values within an interval  $[r_1, r_2]$  is equivalent to taking a vertical slice of the cloud. An illustrative example is shown in Figure 9(a). As shown in the figure, if the correlation is strong, the data in the vertical slice of the plot are also essentially the same data belonging to a horizontal slice corresponding to a latency interval  $[l_1, l_2]$ . So, binning by response time should be expected to also produce a corresponding binning by latency. Stimulus-locked averaging of these data should then present an improved resolution not only for response-locked ERPs but also for variable-latency ERPs. Note that this would happen irrespective of whether the correlation between  $R$  and  $L$  is positive or negative. If the correlation is not very strong,

however, we need to take an analytic approach to understand the effects of binning.

In the previous section, we assumed that a deterministic functional relationship  $g$  between response time  $R$  and latency  $L$  existed, and it was invertible.  $R$  was dependent on  $L$  but  $g$  was totally arbitrary. Here, instead, we want to look at an orthogonal set of assumptions. We will assume that the joint distribution between  $R$  and  $L$  is arbitrary but that there exists a relationship between  $R$  and  $L$  of the form  $R = \varphi(L) + E$ , where  $\varphi$  is an arbitrary function (model) and  $E$  is a stochastic variable with density function  $\varepsilon(e)$ , which is statistically independent from  $L$ . Under these assumptions we find that

$$\begin{aligned} w^{[r_1, r_2]}(l) &= \Pr\{r_1 \leq R < r_2 | l\} \\ &= \Pr\{r_1 \leq \varphi(l) + e < r_2\} \\ &= \Pr\{(r_1 - \varphi(l)) \leq e < (r_2 - \varphi(l))\} \\ &= \int_{r_1 - \varphi(l)}^{r_2 - \varphi(l)} \varepsilon(e) de \end{aligned}$$

In other words, the windowing function  $w^{[r_1, r_2]}(l)$  represents the area subtended by the density function of the error term  $\varepsilon$  in an interval of size  $(r_2 - r_1)$ . The position of this interval on the real axis is determined by  $\varphi(l)$ . The speed at which this interval moves along the real axis as  $l$  varies is determined by  $\varphi'(l)$ . Note that if  $\varphi' = 0$ , i.e., if  $\varphi$  does not depend on  $L$ ,  $\varphi(l) = \text{constant}$  and, so,  $w^{[r_1, r_2]}(l) = \text{constant}$ . This then implies that  $\ell^{[r_1, r_2]}(t) \equiv \ell(t)$ . So, binning does not provide resolution improvements for variable latency waves if  $L$  and  $R$  are uncorrelated.

In general, response-time binning improves the resolution of averaging if  $\ell^{[r_1, r_2]}(l)$  is narrower in the time domain (and, correspondingly, wider in the frequency domain) than  $\ell(l)$ . This, in turn, happens if the windowing function  $w^{[r_1, r_2]}(l)$  is narrower than  $\ell(l)$  in the time domain. Naturally, while this is a likely scenario, we cannot be absolutely certain that this will happen because we don’t know the latency distribution  $\ell(l)$ . However, we know that the narrower the error distribution and the bigger  $|\varphi'(l)|$ , the narrower  $w^{[r_1, r_2]}(l)$ . So, we should expect that, in the presence of strong enough correlations between  $L$  and  $R$ , response time binning will increase the resolution of averaging.

To be more precise, one needs to specialize the analysis to specific forms of  $\varphi$ . So, let us consider a specific case where the regression function of  $R$  on  $L$  is linear, i.e.,  $\varphi(L) = a + bL$  and, so,  $R = a + bL + \varepsilon$ , where  $a$  and  $b$  are the regression coefficients and  $\varepsilon$  is a Gaussian stochastic variable with zero mean and variance  $\sigma^2$ . Under these assumptions, we find that

$$\begin{aligned} w^{[r_1, r_2]}(l) &= \Pr\{(r_1 - a - bl) \leq \varepsilon < (r_2 - a - bl)\} \\ &= \frac{1}{2} \left( \text{erf} \left( \frac{r_2 - a - bl}{\sigma\sqrt{2}} \right) - \text{erf} \left( \frac{r_1 - a - bl}{\sigma\sqrt{2}} \right) \right) \end{aligned}$$

In other words, the windowing function  $w^{[r_1, r_2]}(l)$  represents the area subtended by a Gaussian function over an interval. Note that the condition  $\varphi'(l) = 0$  here corresponds to  $b = 0$ . Standard regression gives  $b = \text{Cov}[R, L] / \text{Var}[L]$ . So, if  $\text{Cov}[R, L] = 0$ , binning gives no resolution benefit for variable-latency waves.

To give an idea of the width and shape of this function in realistic situations, we consider the scatterplot of response times vs latencies of P300 for the ‘accurate condition’ reported in Kutas et al., 1977, p. 794, Figure 2, left. We digitized the figure, and we redraw it in Figure 9(b). The regression line provided with the original figure was  $L = 0.57 \times R + 156$  ms ( $r = 0.66$ ,  $F = 185.65$ ,  $df = 240$ ), which shows a significant correlation between  $R$  and  $L$ . Solving for  $R$ , we obtain  $R = 1.7544 \times L - 273.7$  ms. Using

the data in Figure 9(b), we then estimated  $\sigma \cong 274.7$  ms for this line (as standard,  $\sigma$  was estimated using the RMS of residuals). We then selected the response time bin delimited by  $r_1 = 660$  ms and  $r_2 = 800$  ms, which is indicated by the vertical dashed lines in the figure. With these data in hand, we computed the windowing function  $w^{[r_1, r_2]}(l)$  corresponding to this bin. The function is almost a perfect Gaussian with a mean of 570 ms (indicated by the horizontal dashed line in the figure) and a standard deviation of 160 ms. The function is shown in Figure 9(b) rotated by  $90^\circ$  so that its abscissas correspond to the latency of P300s. The figure also reports a P300 latency histogram (a discretization of the estimated  $\ell(t)$ , again rotated by  $90^\circ$ ). As one can easily see, a significant fraction of the latency histogram does not overlap with the windowing function. Thus,  $\ell^{[r_1, r_2]}(t)$  is narrower than  $\ell(t)$ , resulting in the average over this bin having a significantly higher resolving power in relation to P300s than the ordinary average.

### Resolution of Averages Across Subjects and Grand Averages

The theory developed above makes no assumptions as to whether the trials being averaged relate to a single subject or multiple subjects. Although it is perhaps most naturally applicable in a single-subject setting, in this section, we show that the theory is also valid for averages across subjects and grand averages. What changes in different settings are the response-time and latency-distributions that determine the degree of blurring characterizing each averaging technique.

Let  $\rho_i(t)$  and  $\ell_i(t)$  be the response-time and latency distributions of the  $i$ th subject, respectively. Under the same assumptions as for Equation (3), a subject's stimulus-locked average is:

$$a_i(t) = s(t) + v(t) \star \ell_i(t) + r(t) \star \rho_i(t). \quad (4)$$

Substituting this result in Equations (1) and (2) and simplifying, we can express grand averages and averages across subjects as follows

$$a_g(t) = s(t) + v(t) \star \ell_g(t) + r(t) \star \rho_g(t), \quad (5)$$

$$a_a(t) = s(t) + v(t) \star \ell_a(t) + r(t) \star \rho_a(t), \quad (6)$$

where  $\rho_g(t) = \sum_i \frac{1}{n} \rho_i(t)$  and  $\ell_g(t) = \sum_i \frac{1}{n} \ell_i(t)$  are the means of the single-subject response-time and latency distributions, while  $\rho_a(t) = \sum_i \frac{m_i}{N} \rho_i(t)$  and  $\ell_a(t) = \sum_i \frac{m_i}{N} \ell_i(t)$  are the response-time and latency distributions across subjects, respectively ( $n$ ,  $m_i$ , and  $N$  are defined as before).

Because of natural between-subject variability, we should expect to find that response-time distributions,  $\rho_i(t)$ , and latency distributions,  $\ell_i(t)$ , will vary in shape and location across subjects. Therefore, being  $\rho_g(t)$ ,  $\ell_g(t)$ ,  $\rho_a(t)$ , and  $\ell_a(t)$  weighted sums of such distributions, they will generally be wider than their single-subject counterparts. Thus,  $a_g(t)$  and  $a_a(t)$  will be affected by a "between-subjects" low-pass filtering effect in addition to the "within-subject" blurring characterizing single-subject ERP averages.

Since Equations (5) and (6) have exactly the same form as Equation (3) and their convolution kernels are low-pass filtering ones as those in that equation, the theory developed in the section entitled *Resolution of Variable-latency ERPs* is applicable to grand averages and averages across subjects as it is to single-subject data. Applying response-time binning will, therefore, increase the resolving power also of group averages.

## Conclusions

Stimulus-locked, response-locked, and ERP-locked averaging are the standard methods for reducing artifacts as well as precisely evaluating the shape, amplitude, and latency of specific waves in ERP analysis. All have been exceptionally effective in building up our knowledge on how the brain reacts to stimuli and on the processes that may take place in different tasks.

However, they all suffer from what we could call a key-hole or a magnifying-glass effect. That is, while these techniques are able to increase the resolution of specific ERPs, they do so at the cost of putting everything else out of focus. To build a clearer picture of the ERPs evoked in the brain during a task, an experimenter needs to carefully analyze and qualitatively integrate the averages produced by multiple techniques. This is particularly difficult to do for variable-latency ERPs which are not locked with externally measurable synchronizing events, such as the onset of a stimulus or the response of a participant, because they may effectively fall in the blind spot for both stimulus-locked and response-locked averaging.

Some variable-latency ERPs could be resolved by a suitable ERP-locked averaging process. However, even large and conspicuous waves such as the P300 are difficult to detect on a trial by trial basis. So, averaging based on the latency of waves identified by a detection algorithm may, in fact, lead to mixing the ERPs of interest with other totally unrelated elements, thereby biasing and distorting the result. In addition, ERP-locked averaging requires prior knowledge about the presence and shape of the target wave. In practice, this prevents the use of the method to reveal novel or unsuspected waves.

In this paper, we have proposed an extremely simple technique—binning trials based on response times and then averaging—that can alleviate the problems mentioned above. The technique is based on a simple cognitive homogeneity assumption: that roughly the same cognitive processes and ERPs occur in trials where stimulus condition, participant response, and response time are approximately the same. For this reason, in such trials, the distribution of latencies of all variable-latency ERPs (including those phase-locked with the response) should be narrower than if one considered an undivided dataset. As a result, averaging the trials in a response-time bin should provide a clearer picture of the patterns of brain activity taking place in the conditions associated with those trials.

We assessed the binning technique both empirically and theoretically. For empirical validation we used an experiment in which the task is relatively difficult, requiring identifying and conjoining multiple features, and where response times varied from around 400 ms to over 2 sec. We evaluated the results in a number of ways, including: a comparison between stimulus-locked and response-locked averages, which showed how these are essentially identical under response-time binning; an analysis of statistical significance of inter-bin amplitude differences using Kolmogorov-Smirnov-grams; and an analysis of the signal-to-noise ratios with and without binning. From the theoretical point of view, we provided a comprehensive analysis of the resolution of single-subject averages, grand averages, and averages across subjects, which showed that there are resolution benefits in applying response-time binning even when there is still a substantial variability in the latency of variable-latency ERPs after response-time binning. An improvement of resolving power can be expected whenever there is some correlation (positive or negative) between response time and the latency of an ERP.

This body of evidence suggests that averaging after response-time binning produces clearer representations of brain activity, revealing ERPs and helping in the evaluation of the amplitude and latency of ERP waves. Additionally, the method is extremely simple to use (even retrospectively) and requires no prior knowledge on the ERPs to be enhanced or revealed by the averaging process.

Naturally, the binning method has also limitations. For example, a variety of factors determine whether and how a subject perceives the stimuli and the strategy adopted to decide which answer to produce to such stimuli. These include: the stimuli presented in previous trials, whether the subject's attention or gaze shifted as a result of earlier stimuli, the subject inadvertently zoning out at the time stimulus presentation for the present trial and the corresponding need to resort to guessing, etc. As a result, different processes may be taking place in a subject's brain even within trials characterized by the same stimuli, response, and response-time, thereby violating our cognitive homogeneity assumption. In these cases, bin averages will represent a blend of the ERPs produced by such processes, as ordinary averages would.

Also, it is reasonable to assume that, as task complexity increases, the correlation between the latency of ERPs that are not phase-locked with the response (e.g., associated with intermediate psychological operations) and the response itself will be reduced. Therefore, binning might be unable to reduce the temporal variance of such ERPs and, so, would provide no resolution improvement for them.

As to future research, in a sense, we can think of response-time binning as a spot in the middle ground between single-trial analysis and ordinary averages. In the future, we would like to better explore this middle ground. For example, we would like to

see if binning using gradual membership functions can provide even better reconstruction fidelity (particularly in relation to the Gibbs phenomenon), if setting bin sizes on the basis of the noise in the data may be beneficial to make best use of the available trials, if response-locked and stimulus-locked averages can be jointly used (e.g., in the frequency domain) to further refine the reconstruction of ERPs, if it is possible to integrate the information obtained from different bins into a unified representation of ERPs, if the theory can be extended to cases where the cognitive homogeneity assumptions is violated, etc.

A second line of future research relates to the use of averaging in Brain Computer Interfaces (BCIs) (Farwell & Donchin, 1988; Wolpaw, McFarland, Neat, & Forneris, 1991; Pfurtscheller, Flotzinger, & Kalcher, 1993; Birbaumer, Ghanayim, Hinterberger, Iversen, Kotchoubey et al., 1999; Wolpaw, Birbaumer, Heetderks, McFarland, Peckham et al., 2000). Indeed, ERP averaging is also a key element in many BCIs, and it is precisely from trying to understand its effects in BCI that this work originally emerged. Many BCI systems (e.g., Bostanov, 2004; Rakotomamonjy and Guigue, 2008; Citi, Poli, Cinel, & Sepulveda, 2008) make decisions by *repeatedly* presenting all the stimuli in a set and averaging the corresponding outputs produced by a classifier. If all the steps in the calculation of a classifier's output are linear (and in many cases they are), averaging the outputs of the classifier is equivalent to computing the output produced by it in the presence of an average ERP waveform. In other words, effectively many BCI systems rely on ERP averaging. So, our analysis of the effects of averaging is directly applicable to them. We hope that the response-time binning technique will provide us with a deeper understanding of how users of BCI systems responds to stimuli and of what are the best stimuli for BCI control.

## REFERENCES

- Beauducel, A., & Debener, S. (2003). Misallocation of variance in event-related potentials: Simulation studies on the effects of test power, topography, and baseline-to-peak versus principal component quantifications. *Journal of Neuroscience Methods*, *124*(1), 103–112.
- Birbaumer, N., Ghanayim, N., Hinterberger, T., Iversen, I., Kotchoubey, B., Kbler, A., et al. (1999). A spelling device for the paralysed. *Nature*, *398*(6725), 297–298.
- Bonala, B., Boutros, N. N., & Jansen, B. H. (2008). Target probability affects the likelihood that a P300 will be generated in response to a target stimulus, but not its amplitude. *Psychophysiology*, *45*(1), 93–99.
- Bostanov, V. (2004). BCI competition 2003—data sets Ib and Iib: Feature extraction from event-related brain potentials with the continuous wavelet transform and the *t*-value scalogram. *IEEE Transactions on Bio-Medical Engineering*, *51*(6), 1057–1061.
- Burgess, A., & Gruzelier, J. (1999). Methodological advances in the analysis of event-related desynchronization data: Reliability and robust analysis. In G. Pfurtscheller & F. L. da Silva (Eds.), *Handbook of electroencephalography and clinical neurophysiology, revised series 6* (pp. 139–158). Amsterdam, Elsevier.
- Childers, D. G., Perry, N. W., Fischler, I. A., Boaz, T., & Arroyo, A. A. (1987). Event-related potentials: A critical review of methods for single-trial detection. *Critical Reviews in Biomedical Engineering*, *14*(3), 185–200.
- Citi, L., Poli, R., Cinel, C., & Sepulveda, F. (2008). P300-based BCI mouse with genetically-optimized analogue control. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *16*(1), 51–61.
- Cobb, W. A., & Dawson, G. D. (1960). The latency and form in man of the occipital potentials evoked by bright flashes. *The Journal of Physiology*, *152*, 108–121.
- Dien, J., Spencer, K. M., & Donchin, E. (2003). Localization of the event-related potential novelty response as defined by principal components analysis. *Cognitive Brain Research*, *17*, 637–650.
- Do, K. A., & Kirk, K. (1999). Discriminant analysis of event-related potential curves using smoothed principal components. *Biometrics*, *55*(1), 174–181.
- Donchin, E. (1966). A multivariate approach to the analysis of average evoked potentials. *IEEE Transactions on Bio-Medical Engineering*, *13*(3), 131–139.
- Donchin, E., & Heffley, E. (1978). Multivariate analysis of event-related potential data: A tutorial review. In D. Otto (Ed.), *Multidisciplinary perspectives in event-related brain potential research*, number EPA-600/9-77-043 (pp. 555–572). Washington, DC: U.S. Government Printing Office.
- Donchin, E., & Lindsley, D. B. (Eds.). (1968). *Average evoked potentials: methods, results, and evaluations*, number NASA SP-191, San Francisco: NASA.
- Donchin, E., Ritter, W., & McCallum, W. C. (1978). Cognitive psychophysiology: The endogenous components of the ERP. In E. Callaway, P. Tueting, & S. H. Koslow (Eds.), *Event-related brain potentials in man*. New York: Academic Press.
- Esterman, M., Prinzmetal, W., & Robertson, L. (2004). Categorization influences illusory conjunctions. *Psychonomic Bulletin & Review*, *11*(4), 681–686.
- Farwell, L. A., & Donchin, E. (1988). Talking off the top of your head: Toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and Clinical Neurophysiology*, *70*(6), 510–523.
- Gasser, T., Mcks, J., & Khler, W. (1986). Amplitude probability distribution of noise for flash-evoked potentials and robust response estimates. *IEEE Transactions on Biomedical Engineering*, *33*(6), 579–584.

- Gratton, G., Coles, M. G., Sirevaag, E. J., Eriksen, C. W., & Donchin, E. (1988). Pre- and poststimulus activation of response channels: A psychophysiological analysis. *Journal of Experimental Psychology. Human Perception and Performance*, *14*(3), 331–344.
- Handy, T. C. (Ed.). (2004). *Event-related potentials. A method handbook*. Cambridge, MA: MIT Press.
- Hansen, J. C. (1983). Separation of overlapping waveforms having known temporal distributions. *Journal of Neuroscience Methods*, *9*(2), 127–139.
- Huber, P. J. (1981). *Robust statistics*. New York: John Wiley and Sons.
- Hyvärinen, A., Karhunen, J., & Oja, E. (2001). *Independent component analysis*. New York: John Wiley & Sons.
- Jung, T. P., Makeig, S., Westerfield, M., Townsend, J., Courchesne, E., & Sejnowski, T. J. (2001). Analysis and visualization of single-trial event-related potentials. *Human Brain Mapping*, *14*(3), 166–185.
- Kayser, J., & Tenke, C. E. (2006). Principal components analysis of laplacian waveforms as a generic method for identifying ERP generator patterns: I. Evaluation with auditory oddball tasks. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology*, *117*(2), 348–368.
- Keus, I. M., Jenks, K. M., & Schwarz, W. (2005). Psychophysiological evidence that the SNARC effect has its functional locus in a response selection stage. *Brain Research. Cognitive Brain Research*, *24*(1), 48–56.
- Kok, A. (1988). Overlap between p300 and movement-related-potentials: A response to verleger. *Biological Psychology*, *27*(1), 51–58.
- Kopp, B., Rist, F., & Mattler, U. (1996). N200 in the flanker task as a neurobehavioral tool for investigating executive control. *Psychophysiology*, *33*(3), 282–294.
- Kutas, M., McCarthy, G., & Donchin, E. (1977). Augmenting mental chronometry: The P300 as a measure of stimulus evaluation time. *Science (New York, N.Y.)*, *197*(4305), 792–795.
- Lindsley, D. B. (1968). Average evoked potentials—achievements, failures and prospects. In E. Donchin & D. B. Lindsley (Eds.), *Average evoked potentials: Methods, results, and evaluations*, chapter 1. Washington, DC: NASA.
- Luck, S. J. (2005). *An introduction to the event-related potential technique*. Cambridge, MA: MIT Press.
- Luck, S. J., & Hillyard, S. A. (1990). Electrophysiological evidence for parallel and serial processing during visual search. *Perception & Psychophysics*, *48*(6), 603–617.
- Makeig, S., Bell, A. J., Jung, T.-P., & Sejnowski, T. J. (1996). Independent component analysis of electroencephalographic data. In D. S. Touretzky, M. C. Mozer, & M. E. Hasselmo (Eds.), *Advances in neural information processing systems (volume 8)*, pp. 145–151. Cambridge, MA: MIT Press.
- Makeig, S., Jung, T. P., Bell, A. J., Ghahremani, D., & Sejnowski, T. J. (1997). Blind separation of auditory event-related brain responses into independent components. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(20), 10979–10984.
- Makeig, S., Westerfield, M., Jung, T. P., Covington, J., Townsend, J., Sejnowski, T. J., & Courchesne, E. (1999). Functionally independent components of the late positive event-related potential during visual spatial attention. *The Journal of Neuroscience*, *19*(7), 2665–2680.
- Makeig, S., Westerfield, M., Jung, T. P., Enghoff, S., Townsend, J., Courchesne, E., & Sejnowski, T. J. (2002). Dynamic brain sources of visual evoked responses. *Science (New York, N.Y.)*, *295*(5555), 690–694.
- McCarthy, G., & Donchin, E. (1981). A metric for thought: A comparison of P300 latency and reaction time. *Science (New York, N.Y.)*, *211*(4477), 77–80.
- Nieuwenhuis, S., Yeung, N., & Cohen, J. D. (2004). Stimulus modality, perceptual overlap, and the go/no-go N2. *Psychophysiology*, *41*(1), 157–160.
- Nieuwenhuis, S., Yeung, N., van den Wildenberg, W., & Ridderinkhof, K. R. (2003). Electrophysiological correlates of anterior cingulate function in a go/nogo task: Effects of response conflict and trial type frequency. *Cognitive, Affective, and Behavioral Neuroscience*, *3*(1), 17–26.
- Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, *33*(3), 1065–1076.
- Pfurtscheller, G., Flotzinger, D., & Kalcher, J. (1993). Brain-computer interface: A new communication device for handicapped persons. *Journal of Microcomputer Applications*, *16*(3), 293–299.
- Polich, J., & Comerchero, M. D. (2003). P3a from visual stimuli: Typicality, task, and topography. *Brain Topography*, *15*(3), 141–152.
- Rakotomamonjy, A., & Guigue, V. (2008). BCI competition III: dataset II—ensemble of SVMs for BCI P300 speller. *IEEE Transactions on Bio-medical Engineering*, *55*(3), 1147–1154.
- Roth, W. T., Ford, J. M., & Kopell, B. S. (1978). Long-latency evoked potentials and reaction time. *Psychophysiology*, *15*(1), 17–23.
- Rousselet, G. A., Husk, J. S., Bennett, P. J., & Sekuler, A. B. (2008). Time course and robustness of ERP object and face differences. *Journal of Vision*, *8*(12), 3.1–3.18.
- Salisbury, D. F., O'Donnell, B. F., McCarley, R. W., Nestor, P. G., Faux, S. F., & Smith, R. S. (1994). Parametric manipulations of auditory stimuli differentially affect p3 amplitude in schizophrenics and controls. *Psychophysiology*, *31*(1), 29–36.
- Salisbury, D. F., Rutherford, B., Shenton, M. E., & McCarley, R. W. (2001). Button-pressing affects p300 amplitude and scalp topography. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology*, *112*(9), 1676–1684.
- Schimmel, H. (1967). The (+/-) reference: Accuracy of estimated mean components in average response studies. *Science (New York, N.Y.)*, *157*(784), 92–94.
- Spencer, K. M. (2004). Averaging, detection and classification of single-trial ERPs. In T. C. Handy (Ed.), *Event-related potentials. A method handbook*, chapter 10. Cambridge, MA: MIT Press.
- Spencer, K. M., Abad, E. V., & Donchin, E. (2000). On the search for the neurophysiological manifestation of recollective experience. *Psychophysiology*, *37*(4), 494–506.
- Streeter, D. N., & Raviv, J. (1966). Research on advanced computer methods for biological data processing. *AMRL-TR-66-24 Aerospace Medical Research Laboratories (6570th)* (pp. 1–52).
- Thornton, A. R. D. (2008). Evaluation of a technique to measure latency jitter in event-related potentials. *Journal of Neuroscience Methods*, *168*(1), 248–255.
- Töllner, T., Gramann, K., Müller, H. J., Kiss, M., & Eimer, M. (2008). Electrophysiological markers of visual dimension changes and response changes. *Journal of Experimental Psychology. Human Perception and Performance*, *34*(3), 531–542.
- Verleger, R., Gasser, T., & Mcks, J. (1982). Correction of EOG artifacts in event-related potentials of the EEG: Aspects of reliability and validity. *Psychophysiology*, *19*(4), 472–480.
- Wagner, P., Rschke, J., Grzinger, M., & Mann, K. (2000). A replication study on P300 single trial analysis in schizophrenia: Confirmation of a reduced number of 'true positive' P300 waves. *Journal of Psychiatric Research*, *34*(3), 255–259.
- Wastell, D. G. (1977). Statistical detection of individual evoked responses: An evaluation of Woody's adaptive filter. *Electroencephalography and Clinical Neurophysiology*, *42*(6), 835–839.
- Wolpaw, J. R., Birbaumer, N., Heetderks, W. J., McFarland, D. J., Peckham, P. H., Schalk, G., et al. (2000). Brain-computer interface technology: A review of the first international meeting. *IEEE Transactions on Rehabilitation Engineering*, *8*(2), 164–173.
- Wolpaw, J. R., McFarland, D. J., Neat, G. W., & Forneris, C. A. (1991). An EEG-based brain-computer interface for cursor control. *Electroencephalography and Clinical Neurophysiology*, *78*(3), 252–259.
- Woodman, G. F., & Luck, S. J. (1999). Electrophysiological measurement of rapid shifts of attention during visual search. *Nature*, *400*(6747), 867–869.
- Woody, C. D. (1967). Characterization of an adaptive filter for the signal analysis of variable latency neuroelectric signals. *Medical and Biological Engineering and Computing*, *5*, 539–554.
- Yin, G., Zhang, J., Tian, Y., & Yao, D. (2009). A multi-component decomposition algorithm for event-related potentials. *Journal of Neuroscience Methods*, *178*(1), 219–227.
- Zhang, J. (1998). Decomposing stimulus and response component waveforms in ERP. *Journal of Neuroscience Methods*, *80*(1), 49–63.

(RECEIVED March 27, 2009; ACCEPTED July 1, 2009)