Physiology & Behavior 130 (2014) 13–22

Contents lists available at ScienceDirect

Physiology & Behavior

journal homepage: www.elsevier.com/locate/phb

Reliability of event-related potentials: The influence of number of trials and electrodes ☆

Renske Huffmeijer *, Marian J. Bakermans-Kranenburg, Lenneke R.A. Alink, Marinus H. van IJzendoorn

Centre for Child and Family Studies, Leiden University, P.O. Box 9555, 2300 RB Leiden, The Netherlands
Leiden Institute for Brain and Cognition (LIBC), Leiden University, P.O. Box 9600, 2300 RC Leiden, The Netherlands

HIGHLIGHTS

- ERP amplitudes generally showed adequate to excellent test–retest reliability.
- Averaging across several electrodes or trials improved reliability of P3 amplitude.
- We recommend including at least 30 trials for early, narrowly distributed components.
- Substantially more trials are needed for later, broadly distributed components.

ARTICLE INFO

Article history:
Received 27 September 2013
Received in revised form 27 January 2014
Accepted 7 March 2014
Available online 16 March 2014

Keywords:
Event-related potentials
Retest reliability
Emotional facial expressions
Feedback

ABSTRACT

The reliability of event-related potentials (ERPs) is an important factor determining the value of studies relating ERP components to individual differences. However, studies examining the reliability of ERPs are surprisingly scarce. The current study examines the test–retest reliability of ERP components (VPP, N170, MFN, FRN, P3, and LPP) in response to feedback stimuli combining performance feedback with emotional faces in a sample of healthy female adults. In general, ERP amplitudes showed adequate to excellent test-retest reliability across a 4-week interval, depending on the component studied. Averaging ERP amplitudes across several electrodes yielded more reliable measurements than relying on a single electrode. Averaging across multiple trials substantially improved reliability. We recommend including at least 30 trials for early, spatio-temporally narrowly distributed components (such as VPP), but substantially more, at least 60 trials, for later, broadly distributed components such as the P3.

1. Introduction

Throughout the scientific literature event-related potentials (ERPs) are widely used to study aspects of information processing inside the brain. Many studies focus on individual differences, comparing ERP amplitudes and latencies between (sub)groups of participants characterized by some psychological, behavioral, or other trait, or correlating ERPs with measures of such individual characteristics. The value of the insights these studies offer depends to a large extent on the reliability and validity of the ERP measures. To adequately evaluate group differences in or correlations with electrocortical activity, sound measures of electrocortical activity are essential. A sound measure is both reliable (repeatable; repeated measurement of the same characteristic in the same individual under the same conditions must give the same outcome) and valid (a measure must measure [only] the characteristic it is intended to measure). In fact, reliability is considered to be a necessary (though not sufficient) condition for validity (e.g., [4]). There are a number of factors, including a low signal-to-noise ratio and trial-to-trial variation in the amplitude and phase of the underlying neural activity (cf., [3,29]), that may increase measurement error and as a result compromise the reliability of ERP measures. Reliability is thus a serious concern for ERP research, and accurate estimates of ERP reliability are essential to evaluate the trustworthiness of study results. Given the importance of reliable (and valid) measures, studies examining the reliability of ERP measures are surprisingly scarce.

Only a few previous studies have examined the reliability of amplitude and/or latency measures of ERP components. The best studied ERP components are perhaps the auditory mismatch negativity (MMN) and components reflecting error-related brain activity (error-related

☆ MJB-K and MHvIJ were supported by research awards from the Netherlands Organisation for Scientific Research (MHvIJ: NWO SPINOZA prize; MJBK: VICI grant no. 453-09-003), by the Gravitation program of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organisation for Scientific Research (NWO grant number 024.001.003).

* Corresponding author at: Centre for Child and Family Studies, Leiden University, P.O. Box 9555, 2300 RB Leiden, The Netherlands. Tel.: +31 71 5274080; fax: +31 71 5273945.
E-mail address: rhuffmeijer@fsw.leidenuniv.nl (R. Huffmeijer).

http://dx.doi.org/10.1016/j.physbeh.2014.03.008
0031-9384/

negativity [ERN], error-positivity [Pe], and feedback-related negativity [FRN]). MMN, ERN, Pe, and FRN amplitudes have been found to possess moderate to good reliability, with reported (test–retest or split-half) correlations usually exceeding .50 (e.g., [8,13,22,27,30,35,38]). The reliability of amplitude measures of the P3 component has also been examined in several studies. In these studies reliability was concluded to be at least moderate (e.g., [13,19,31]). Of note is that, although some studies have obtained adequate reliabilities for ERP latency measures (e.g., [32]), lower reliability is often reported for latency than for amplitude measures (e.g., [27,37,38]). In addition, the reliability of ERP measures may be affected by the number of trials included in the averaged ERP [25].

The current study examines the test–retest reliability of several ERP components in response to salient visual stimuli: emotional faces with performance feedback. Similar stimuli have been used in several studies (e.g., [15,16,36]) and results are thus directly relevant to research practice. We focus on amplitude measures of six components: Vertex Positive Potential (VPP), N170, Medial Frontal Negativity (MFN), difference FRN, P3, and Late Positive Potential (LPP). Because the stimuli convey both performance feedback and facial emotions, the current study focuses on ERP components related to processing of both feedback (MFN/FRN) and faces (VPP, N170) as well as more general processes (P3, LPP). Some of these components (MFN/FRN and P3) have been the subject of previous studies of ERP reliability, whereas the test–retest reliability of other components (VPP, N170, LPP) has, to the best of our knowledge, not been studied previously. We expect reliabilities of all components to be moderate to good (in accordance with results of previous studies), but to be dependent on the number of trials included in the averaged ERP and the number of electrodes used to quantify ERP amplitude.

The VPP, a positive deflection in the ERP that peaks at frontocentral electrode sites approximately 140–180 ms after stimulus onset, and N170, a negative going occipitotemporal right hemisphere dominant component in the same time-range, are thought to represent two sides of the same generator dipoles in occipitotemporal cortex [17]. Both components have been associated with the configural processing of faces, and usually show larger amplitudes in response to emotional compared to neutral expressions [23,33]. If differences in the reliability of VPP and N170 amplitudes are observed, this may be an argument to favor one component over the other in studies of face processing.

The MFN, a frontocentrally maximal negative-going ERP component, peaking approximately 250 ms after the onset of a feedback stimulus is often used to quantify the feedback-related negativity (FRN). These components have been associated with performance monitoring and are more negative after negative outcomes compared to positive outcomes [2,12,24,26]. In the current study, both the amplitude of the 'raw' MFN in response to positive and negative feedback, as well as the amplitude of a difference FRN will be obtained. This FRN is the difference in amplitude between MFN responses to positive and negative feedback. The effect of performance feedback on MFN/FRN amplitude has been suggested to depend on the unpredictability of the feedback [14]. MFN/FRN amplitude may thus be a reliable index of feedback processing only if the feedback is relatively unpredictable, i.e., when participants actually need to process the feedback stimulus to learn whether a response was correct. In the current study feedback stimuli are presented during performance of a flanker task. When performing this task participants generally know whether a response was correct before feedback was presented (i.e., feedback is predictable). However, we used a version of the flanker task in which responses could also be late (i.e., exceed a deadline for correct responses). Because it is difficult for participants to estimate whether a response was on-time or late, participants must rely on the feedback for this information (i.e., late feedback is unpredictable). Because error feedback can be predicted, but late feedback is relatively unpredictable, we expect that reliable MFN and FRN amplitudes will be generated in response to late feedback only.

The P3 (or P300) is a broadly, parietally distributed component that peaks between 300 and 500 ms after stimulus onset [11,34]. The LPP is a centroparietally distributed, positive-going modulation of the ERP beginning about 300–400 ms after stimulus onset [10,28]. Both components are thought to reflect attentional orienting to (motivationally or emotionally) salient stimuli [9,10,28]. In fact, the two components may reflect partially the same processes and neural activity [11]. We will investigate whether both components are reliably generated.

Because evidence suggests that the number of trials included in the averaged ERP may affect reliability [25], we will investigate effects of including additional trials in the averaged ERP on the reliability of VPP and P3 amplitudes. We choose to study VPP and P3 amplitudes because these components differ markedly in their temporal and spatial distributions as well as in the processes that generate them. The VPP shows a relatively limited distribution, both in time and space, and is thought to reflect an early, relatively automatic stage of processing (processing face configuration; [23]). The P3 shows a broad temporal and spatial distribution, and is thought to be sensitive to more controlled (as well as automatic) attentional processes [11,34]. We expect that reliability will improve when more trials are added, and we search for an optimal number of trials. Lastly, we investigate whether averaging across multiple electrodes improves reliability of a broadly distributed component (P3). We expect that averaging across multiple electrodes will improve reliability, as long as the electrodes adequately cover the component's scalp distribution, by reducing the influence of random noise on any single electrode.

2. Materials and method

2.1. Participants

A total of 12 female undergraduate students, aged 18–22 years ($M = 19.17$, $SD = 1.12$), took part in the ERP experiment that consisted of two sessions separated by approximately four weeks. One participant completed only one session, and data of another participant could not be analyzed because of a very low error rate. The final sample thus consisted of 10 female undergraduate students (age: $M = 19.20$, $SD = 1.14$). They were paid 40 Euros or received course credits for participation. Exclusion criteria included colorblindness, smoking, alcohol and drug abuse, neurological and psychiatric disorders, pregnancy, breastfeeding, and use of medication (except oral contraceptives). The study was approved by the ethics committee of the Leiden University Medical Center.

2.2. Procedure

Participants were asked to come to our laboratory for two experimental sessions, separated by approximately four weeks. Informed consent was obtained at the beginning of the first session. Participants were fitted with an electrode net after which they completed the flanker task (with a short break after the fourth block). At the start of each session, halfway through and after completion of the task participants completed some questionnaires. Data regarding these questionnaires will be presented elsewhere.

2.3. Experimental tasks

During each session, participants completed eight 72-trial blocks of a modified Eriksen flanker task [7], preceded by a 72-trial practice block. Target stimuli consisted of a row of five arrows (7.4° × 1.4° visual angle), presented for 50 ms, all pointing in the same direction (congruent targets), or with the middle arrow pointing in the opposite direction (incongruent targets). Target stimuli were preceded by a fixation cross, presented in black for 1000 ms and then in red for 800–1200 ms (to draw attention to the screen). The participants