Internal Consistency in Event-Related Potentials associated with the Eriksen Flanker Experiment

Indy Bernoster[1,*],
Supervisor: Prof. dr. Patrick J.F. Groenen,
Co-reader: Prof. dr. Dennis Fok

Master Thesis
Erasmus University Rotterdam

Abstract

In the present psychiatric literature, average Event-Related Potentials (ERPs) are often linked to psychiatric disorders. After awareness of making an error, the brain will react. Average ERPs are derived by averaging these brain reactions over all errors. However, if only a few errors are available, the average ERP can be unreliable. Recent studies show attention for examining the reliability of ERPs. These studies try to find the number of errors that makes ERPs reliable, or, internal consistent. Olvet and Hajcak (2009), Marco-Pallares et al. (2011), Pontifex et al. (2010), Meyer et al. (2013), Rietdijk et al. (2014) use Cronbach's α to measure internal consistency. However, this gives two problems. First, of all errors made by a participant, only some errors are used to obtain α, while taking another set of errors could give another value for α. Secondly, one of the main assumptions underlying Cronbach's α is violated. This assumption states that precisely the same trials need to be used over participants. Nevertheless, it is quite unlikely that participants fail exactly the same trials. The main goal of this research is to investigate whether these problems bias the number of errors that these studies find. Furthermore, having reliable average ERPs depends on whether brain activity is independent of error trials. Therefore, another goal is to justify averaging over all error trials, that is, to show independency of brain activity over error trials. To meet the goals, we examine a random parameter model, empirical distributions of Cronbach's α, and a simulation study. Data containing brain activity values of 37 participants in a Eriksen Flanker experiment is used. It turns out that ERPs are independent of error trials and that taking a specific set of errors to compute α can significantly bias the number of errors needed for internal consistency.

Keywords: Cronbach's α, Internal Consistency, ERPs, Simulation

*Corresponding author
[1]Student number: 349277, email address: indybernoster@outlook.com

Contents

1. Introduction

In the field of psychiatry (Tantam, 2000, Wheaton, 1980), brain activity is often related to psychiatric disorders. This brain activity is measured in terms of Event-Related Potentials (ERPs), which consist of Error-Related Negativity (ERN) and Error Positivity (Pe). The ERN and Pe values are only observed with consciousness of making an error. For example, consider a simple task where one needs to press the left arrow on a keyboard if a left arrow is shown on a screen and the right arrow on a keyboard if a right arrow is shown. Now, the response can be either correct or false, where a false response results in an error. If one becomes aware of making an error, the brain will react. It is from this reaction that the ERN and Pe values are derived. In the present literature, higher ERN values are, for instance, related to obsessive-compulsive disorder (Gehring et al., 2000, Johannes et al., 2001), anxiety (Ladouceur et al., 2006) and depression (Chiu and Deldin, 2007, Holmes and Pizzagalli, 2008).

Usually, these kind of studies consider participants who perform a large number of simple tasks, to be called trials in the following. Often, participants fail several of these trials, resulting in a number of error trials per participant for which ERN and Pe values are available. Then, the average of these ERPs is derived by averaging ERPs over all error trials. It is this average that is linked to psychiatric disorders. However, the ERN or Pe value, coming from one specific error, could be partly subjected to coincidence. Therefore, the average ERP for a participant with few errors can be unreliable. So, as ERPs are only observed with errors, there need to be enough error trials to infer reliable conclusions from the brain activity. Therefore, several studies investigate the reliability of ERPs (Olvet and Hajcak, 2009, Marco-Pallares et al., 2011, Pontifex et al., 2010, Meyer et al., 2013, Rietdijk et al., 2014). This reliability is analysed in terms of internal consistency, where internal consistency is defined as the similarity of brain activity across error trials (Wöstmann et al., 2013). A well-known measure for internal consistency, used by all these studies, is Cronbach's α. These reliability studies examine the number of errors that result in internal consistent ERPs. This number of errors is used as a selection criterion. Namely, participants with too few errors for having a reliable average ERP are eliminated from the sample. For the remaining participants, average ERPs are derived by averaging all error trials.

To decide upon the number of errors used for the selection criterion, the current reliability studies use the following procedure. All participants with less than fourteen errors are excluded, such that Cronbach's α can be computed for the first two, four, six, eight, ten, twelve, and fourteen errors. Now, the number of errors needed for internal consistent ERPs can be determined from these values for Cronbach's α. Namely, if α is high enough, the corresponding number of errors will be sufficient for internal consistency. In this way, Olvet and Hajcak (2009) conclude that the number of error trials necessary for a stable ERN or Pe is between 6 and 10 for ERN and between 2 and 6 for Pe. The other studies show comparable results. Let us call the approach to compute Cronbach's α in these studies the OH-method.

Unfortunately, the OH-method contains two problems. Firstly, the first fourteen error trials are used to compute Cronbach's α, while other errors are ignored. Consequently, only one specific Cronbach's α is selected out of a whole set of α's that can be obtained by choosing different combinations of error trials. Secondly, participants do not necessarily fail the same error trials. However, computing Cronbach's α implies that it is assumed that the first error, the second error, and so on, are made on the same trial, which is highly unlikely. Accordingly, this method ignores the definition of Cronbach's α, namely, the method ignores that α need to be computed over the same trials. As a consequence, Cronbach's α cannot be interpreted as a lower bound of the reliability of the ERPs.

In addition to these problems, averaging over all error trials (instead of a selection of error trials as, for example, determined by Cronbach's α) can also influence reliability of the average ERPs. For the averaging over all errors to be justified, the ERPs within a participant should be independent over error trials. For example, suppose that ERPs decrease over time as making the first error will be more impactful than making the last one and consider two participants that are exactly the same except for one making many errors and the other making just a few errors. Then, averaging all error trials will result in a low average for the participant with many errors and a high average for the one with only a few errors. In this way, we would conclude different brain activity values for the participants, while, in fact, brain activity is the same. Therefore, taking an average over all error trials could be unreliable. Also, it can be the case that different types of trials will give different brain activity values. Here, similar problems can arise. Therefore, we examine whether brain activity is independent across error trials. Investigation of a selection criterion, that is, examining the OH-method, is useful only if averaging over all error trials is justified.

The main purpose of this research is to investigate whether the problems arising from the OH-method bias the number of errors obtained from this method. The corresponding research question is 'Is the number of errors needed for internal consistency of the ERPs coming from the OH-method biased by the problems arising from this method?' If there is bias, we try to find other ways to decide upon the number of errors needed for internal consistent ERPs. An additional goal, that will be explored first because of reasons described above, is to examine whether averaging over all error trials is justified. Therefore, questions as 'Do different types of trials give different brain activity values?' and 'Do ERPs change over time?' need to be answered.

To answer these research questions, different approaches are used. First, we will examine the justification of averaging all trials. To do so, we test for a trend in ERPs and we set up a random parameter model to test whether different types of trials give different brain activity values. Secondly, we investigate whether the number of errors needed for internal consistency is biased by the OH-method. This is done by creating empirical distributions for Cronbach's α and by exploring a simulation study. The empirical distributions are used to decide whether taking the first fourteen errors, instead of a random selection of

errors, gives biased results. The simulation study is used to mimic perfect data and compute Cronbach's α in the right way, so that it can be compared to α coming from the OH-method. Finally, other ways for deciding upon the internal consistency of the ERN or Pe values are considered.

The outline of this research is as follows. We start by explaining important concepts used in this report. Next, we will explain the research goals in more detail. Then, the particular data set used for justification of averaging all error trials and for creating empirical distributions is discussed. Subsequently, we focus on solving the research questions by using multiple approaches. First, we will try to justify taking the average over all error trials. For this purpose a random parameter model is set up. Then, we compute α following the OH-method and compare this value with empirical distributions of Cronbach's α. Finally, we perform a simulation study to examine violation of the assumptions of Cronbach's α. In addition, we give recommendations about finding the number of errors needed for internal consistency in future research. Finally, we conclude and discuss.

2. Concepts

In this section, different concepts necessary for understanding the remainder of this research will be explained. We discuss EEG, ERPs, the Eriksen Flanker experiment and Cronbach's α. In all explanations, we will focus on the details important for this report.

2.1. Electroencephalography

The context of this research comes from experiments done with electroencephalography (EEG). Using EEG, one can measure the electrical activity along the scalp in terms of voltage fluctuation. Someone undergoing an EEG experiment has electrodes placed on several spots on the head. The data for this particular research is based on 32 electrodes placed on the scalp. Figure 1 shows a schematic view of an EEG experiment on the left side. A participant has a special cap on his or her head. Electrodes are connected with the cap and with a computer. Now, electrical activity of the outermost layer of the brain is measured for all shown electrodes in the right part of the figure.

The midline electrodes appear on the line from the nose to the back (coded as Fz, FCz, Cz, CPz, Pz, see the right panel of Figure 1). These electrodes give the most accurate proxies of brain activity, and are widely accepted in the field of cognitive neuroscience (see, for example, Olvet and Hajcak, 2009, Marco-Pallares et al., 2011, Hajcak, 2012). In fact, to our knowledge, all studies usually report only data of midline electrodes. In this study, we will focus on two of these electrodes, namely FCz and Pz.

During an EEG experiment, participants usually perform a large amount of simple trials, for example, pressing the left arrow on a keyboard if a left arrow in shown in the screen and the right arrow if the right arrow is shown. For each participant, EEG measures brain activity for all electrodes during the

experiment on a continuous time scale. Therefore, EEG experiments create a lot of data.

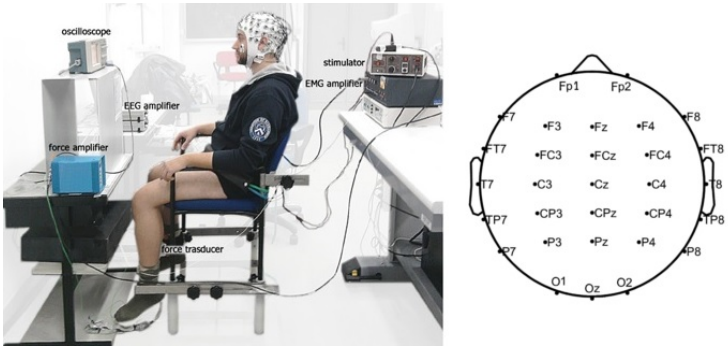


Figure 1: A schematic view of an EEG experiment (left). The 32 electrodes placed on the scalp (right).

As respondents can move while being recorded, brain activity can sometimes contain measurement errors. The Signal-to-Noise ratio (SNR) measures whether brain activity is signal or noise. If the SNR is higher than one, there is more signal, and if it is lower than one, there is more noise.

2.2. Event-Related Potential

As already mentioned, participants in EEG experiments face a large number of simple trials. They can either do a trial correctly or make an error. When a participant becomes aware of making an error, his or her particular EEG data shows different brain activity patterns than when the participant thinks to be correct. An example of two brain activity patterns for the same person is given in Figure 2.

The red line corresponds to an error of this participant, and the black line to a correct answer. This brain activity (in micro Volt) comes from one specific electrode, namely FCz (see Figure 1). The dotted vertical line indicates the moment of answering the trial by pressing a button. After this dotted line, the two brain activity patterns start to differ from each other. The error signal has a high peak followed by a low trough, while the non-error signal roughly has the same level over the entire time span. The brain activity is measured from 100 milliseconds before the moment of answering until 600 seconds after answering, which results in an interval of 700 milliseconds of brain activity.

Using EEG, one obtains brain activity for all electrodes and all trials of each participant. From the brain activity of the error trials the Error-Related Negativity (ERN) and the Error Positivity (Pe) can be decided, which are both examples of an Event-Related Potential (ERP). The ERN can best be measured from the FCz electrode. We quantify ERN using an area measure of the period associated with 25 to 75 milliseconds after the error is made (see Figure 2). The Pe is measured in the same way, but it turns out that the best electrode for

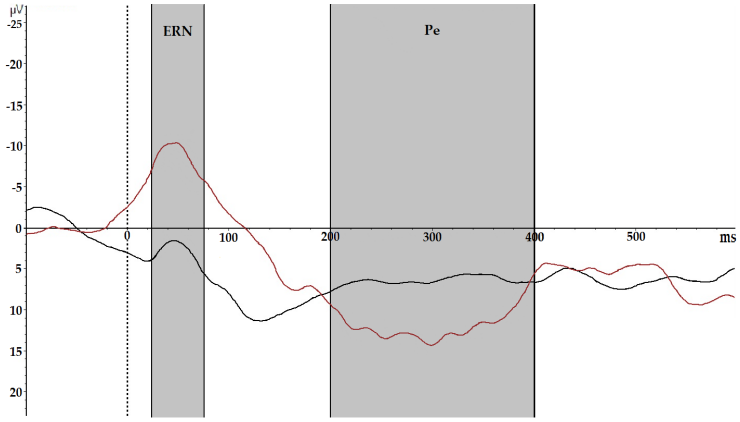


Figure 2: An example of EEG output, showing an error response (red line) and a non-error response (black line), for 700 milliseconds (ms) in micro Volt (μV). The figure shows brain activity for the FCz electrode for one particular participant.

computing Pe is Pz and that this should be done in 200 till 400 milliseconds after the error is made. Note that Figure 2 only shows the FCz electrode, so that the Pe value from this figure will probably be less accurate than the Pe value from the Pz electrode.

2.3. Eriksen Flanker Task

In the Eriksen Flanker experiment (Eriksen and Eriksen, 1974), participants react to a large set of simple Eriksen Flanker tasks. In this study, the tasks consist of identifying characters 'H' and 'S'.

In each task, participants are first shown the upper display in Figure 3. This allows the participant to focus on this point, as the focal letter will be represented here. Then, a black screen follows for a short period. After that, one of the four middle displays in Figure 3 is shown. The middle character (above the red arrow) is the focal letter. Note that the red arrow is not shown to the participants. If the letter on the focal point is an 'S', the participant should respond by pressing '1' on a keyboard. If the character is an 'H', the participant should press '5'. After the participant pressed the button, a feedback display is shown. If the participant responded correct, he or she will see 'OOO' and if the participant responded incorrect, he or she will see 'XXX'. Sometimes, participants do not react, which results in '!'. The feedback displays are shown in the lower part of Figure 3.

The focal letter is flanked by other letters to distract the participant. In this way, as we already saw in Figure 3, we obtain four different stimuli. These can either be congruent, meaning that the distraction letters are the same as the focal letter, or be incongruent, meaning that distraction letters differ from the focal letter. Table 1 represents the four stimuli with their characteristics, that is, the correct key to press on the keyboard and whether the task is congruent or incongruent.

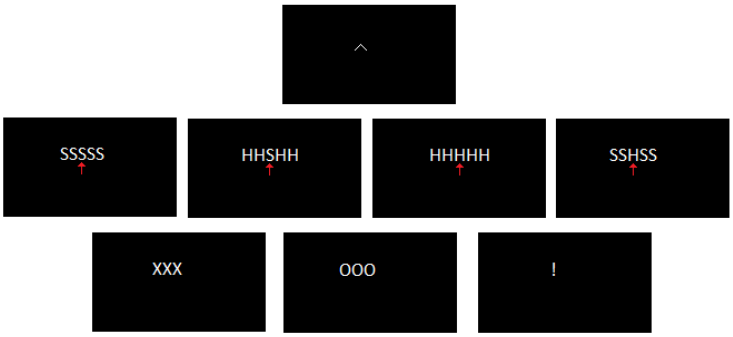


Figure 3: The upper display shows the focal point display. The four middle displays show the stimuli that are shown to participants. The three lower ones show feedback displays.

Table 1: The four different stimuli in the Eriksen Flanker experiment with their characteristics.

Stimuli	Correct	Category
SSSSS	1	Congruent
HHSHH	1	Incongruent
HHHHH	5	Congruent
SSHSS	5	Incongruent

In Eriksen Flanker experiment for this research, interest lies in the EEG response of individuals. EEG participants are settled as in Figure 1. Then, a screen shows them one of the four stimuli and the participant can respond using the keyboard. The ERN and Pe are measured directly after a respondent presses a key. So, for measuring ERN and Pe values, we do not wait for the feedback displays. The reason for this is that participants will already feel whether they pressed the correct or wrong key and thus the brain will react directly. Remark that we can only quantify ERN or Pe if an error is made.

2.4. Cronbach's α

Cronbach's α (Cronbach, 1951) is a measure of internal consistency for a number of items. It measures to what extent those items measure the same phenomena. For example, if we consider a questionnaire, α gives an idea to what extend the questions measure the same concept. Cronbach's α is a lower bound of the reliability of the items (see Appendix A for a derivation). The reliability of the items represents whether measurement errors influence the answers to the items. The higher the reliability, the less answers are influenced by measurement error. The problem with this reliability is that it is based on theoretical quantities, so that we cannot compute it. However, one can calculate Cronbach's α, such that we have a lower bound for this reliability.

Suppose that we have values, on e.g. a questionnaire, y_{ij}, where $i = 1, ..., N$ denotes a participant and $j = 1, ..., K$ denotes a question. Furthermore, define t_i as the sum of the measures over the questions of participant i, that is, $t_i = \sum_{j=1}^{K} y_{ij}$. Now, Cronbach's α is computed (see Cronbach, 1951) as

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{j=1}^{K} s_j^2}{s_t^2}\right), \tag{1}$$

where s_j^2 is the sample variance of the j^{th} question over all respondents and s_t^2 is the sample variance of the total scores (t_i) over all respondents.

Cronbach's α is defined such that the values used to compute the variance of the j^{th} item, all actually correspond to that item. Consequently, the variance s_j^2 does not allow to use values y_{il} with $l \neq j$ as is done in Olvet and Hajcak (2009). Therefore, a main assumption of Cronbach's α is that it is computed over the same items.

Cronbach's α can take values from $-\infty$ till 1 (see Appendix A). To associate Cronbach's α with internal consistency, the relation in Table 2 can be used.

Table 2: The relation between internal consistency and Cronbach's α.

Internal consistency	Cronbach's α
Excellent	$\alpha > 0.9$
High	$0.7 < \alpha \leq 0.9$
Moderate	$0.5 < \alpha \leq 0.7$
Low	$\alpha \leq 0.5$

3. Problem Definition

This section is meant to clarify the research goals. First, we will exactly explain the OH-method, which is used to find the number of errors needed to have internal consistent ERPs. Then, the problems arising from this method will be set out. Next, we explain how averaging over all error trials can also influence the reliability of average ERPs. Finally, we will focus on possible approaches to examine justification of averaging over all errors and to investigate whether the OH-method gives bias in the number of errors needed for internal consistency.

3.1. OH-method

Let y_{ij} be the ERN or Pe for participant i and trial j, where $i = 1, ..., N$ and $j = 1, ..., K$. As we are only interested in the error responses, define

$$x_{ij} = \begin{cases} y_{ij} & \text{if the response is false,} \\ 0 & \text{if the response is correct.} \end{cases}$$

Furthermore, define X as the $N \times K$ matrix for either ERN or Pe, containing the values x_{ij}. Usually, K is high and participants do not make many errors, such that these matrices will be very sparse.

The left matrix of Table 3 shows a hypothetical example of a matrix X, where we have $N = 5$ participants and $K = 6$ trials. This matrix thus contains ERN or Pe values on error trials. The first participant, corresponding to the first row, made errors in trials 1, 2, and 5, with brain activity values x_{11}, x_{12}, and x_{15}, respectively. Furthermore, the left matrix of Table 4 shows the same matrix, but now with the corresponding stimuli instead of the ERN or Pe value. Thus, the first participant failed on stimulus 3, 4, and 2, which correspond to 'HHHHH', 'SSHSS', and 'HHSHH', respectively. Note that, the order of the stimuli is randomised over the respondents, such that each participant faces a different ordering of stimuli. In sum, the left matrices of Tables 3 and 4 form the raw data coming from the Eriksen Flanker experiment.

Table 3: An example of the matrix X in the raw version (left) and as used by the OH-method (right).

$$\begin{bmatrix} x_{11} & x_{12} & 0 & 0 & x_{15} & 0 \\ x_{21} & 0 & x_{23} & x_{24} & 0 & 0 \\ 0 & x_{32} & 0 & x_{34} & x_{35} & x_{36} \\ 0 & 0 & x_{43} & 0 & x_{45} & 0 \\ 0 & 0 & x_{53} & 0 & x_{55} & x_{56} \end{bmatrix} \qquad \begin{bmatrix} x_{11} & x_{12} & x_{15} & 0 & 0 & 0 \\ x_{21} & x_{23} & x_{24} & 0 & 0 & 0 \\ x_{32} & x_{34} & x_{35} & x_{36} & 0 & 0 \\ x_{43} & x_{45} & 0 & 0 & 0 & 0 \\ x_{53} & x_{55} & x_{56} & 0 & 0 & 0 \end{bmatrix}$$

Table 4: The stimuli corresponding to the matrices in Table 3. Note that, 1 = 'SSSSS', 2 = 'HHSHH', 3 = 'HHHHH' and 4 = 'SSHSS'.

$$\begin{bmatrix} 3 & 4 & 0 & 0 & 2 & 0 \\ 1 & 0 & 3 & 3 & 0 & 0 \\ 0 & 4 & 0 & 4 & 1 & 4 \\ 0 & 0 & 3 & 0 & 4 & 0 \\ 0 & 0 & 1 & 0 & 3 & 2 \end{bmatrix} \qquad \begin{bmatrix} 3 & 4 & 2 & 0 & 0 & 0 \\ 1 & 3 & 3 & 0 & 0 & 0 \\ 4 & 4 & 1 & 4 & 0 & 0 \\ 3 & 4 & 0 & 0 & 0 & 0 \\ 1 & 3 & 2 & 0 & 0 & 0 \end{bmatrix}$$

Using these matrices, the OH-method can be explained. Remember that computation of Cronbach's α is defined over the same trials, so in order to compute Cronbach's α, the $N \times K$ matrix X needs to have columns which do not contain any missings, that is, correct answers, or, zeros. Hence, as can be seen from the example, the raw matrices for ERN and Pe are not sufficient as they contain many missings. The OH-method creates usable matrices from those raw matrices so that Cronbach's α still can be computed. To do so, all participants with less than K_m errors are excluded from the data matrix. Then, the ERN or Pe values in the raw matrices (all nonzeros) are pushed to the left. In this way, there is at least a $N \times K_m$ matrix containing columns without zeros. From this matrix, the OH-method computes Cronbach's α over the first k columns, or, over the first k error trials, where $k \in \{2, 4, ..., K_m\}$. Note that,

Olvet and Hajcak (2009) uses $K_m = 14$. Now, α can be compared over the different values for k. The number of errors (k^*) that makes α high enough (following Table 2) is needed to have a reliable measure of the ERPs.

To show the OH-method in the example, consider the right matrices in Tables 3 and 4. Transformation from the left matrices to the right matrices includes pushing the values in the rows to the left as much as possible. Now, Cronbach's α can be computed from the black square, because the columns in this black square do not contain zeros. Note that, in this hypothetical example, $K_m = 2$.

3.2. Problems Arising from the OH-method

Although the OH-method creates a matrix from which Cronbach's α can be computed, it also has some complications. The first problem is that the OH-method violates the main assumption of Cronbach's α as explained in Section 2.4, namely, α should be computed over the same trials. We can define three (sub)assumptions that constitute to being the same trials:

- $\mathcal{A}.1$: Having the same stimulus: brain activity values in the same column should have the same stimulus.
- $\mathcal{A}.2$: Having the same trial number (j): brain activity values in the same column should be the same trial, for example, they are both the tenth trial.
- $\mathcal{A}.3$: Having the same error number: brain activity values in the same column should be the same error, for example, they are both the second error that was made.

From those assumptions, the first two, $\mathcal{A}.1$ and $\mathcal{A}.2$, are violated by the OH-method. To show why, consider again the example of Tables 3 and 4. The fact that the first assumption ($\mathcal{A}.1$) is not met has to do with the randomisation of the stimuli (red square) and the matrix transformation as explained above (green squares). From the red square, we see that although both the first and second participant failed the first trial, they still failed a different stimulus due to randomisation of the stimuli. Namely, the first participant failed on 'HHHHH', while the second failed on 'SSSSS'. From the green squares, we see that in the raw matrices those values are both in the second trial and that they have the same stimulus ('SSHSS'). However, in the transformed matrices, the value for the first participant is still in the second column, while the value for the third participant is placed in the first column. Therefore, those values were correctly placed, but because of transformation they do not match anymore.

The violation of $\mathcal{A}.2$ can be seen from the blue squares. It can, by accidence, happen that the transformed matrix results in the same stimuli in a column. However, those trials are actually different in the left matrix as the third participant made an error in the fourth trial (so trial number (j) is 4) and the fourth participant failed the fifth trial (trial number (j) is 5). However, both trials were 'SSHSS' and both were the second error for the participant. In this way,

those trials came in the same column, but actually do not have the same trial number.

Note that, $\mathcal{A}.3$ is met for the transformed matrix. Namely, the first column contains all errors that are first made by students, the second column contains errors that are made secondly, and so on.

The consequence of violating $\mathcal{A}.1$ and $\mathcal{A}.2$ can be seen very clearly from the hypothetical example. In the right matrices of Tables 3 and 4, the black squares show that the trials in the columns do not match at all, that is, not on stimulus and not on trial number. This will also be the case in the $N \times K$ data matrix X. In this way, Cronbach's α is not computed over the same trials, such that it violates its main assumption and thus cannot be interpreted as the lower bound of the reliability of the ERPs. Therefore, the number of errors coming from this α, that is, k^*, can be biased.

The second problem of the OH-method is that only the first K_m errors are considered, such that, speaking in terms of the example, brain activity values x_{15}, x_{24}, x_{35}, x_{36} and x_{56} are fully ignored. However, Olvet and Hajcak (2009) have three reasons for considering the first K_m error trials. Firstly, taking the first K_m errors compensates violation of the assumption. Namely, taking the first K_m errors will ensure that $\mathcal{A}.3$ is fulfilled. Secondly, Olvet and Hajcak (2009) included different randomisations of error trials in computing α, but did not find other values for α, so that taking the first K_m errors would be justified. A third reason for taking the first K_m errors is that the Signal-To-Noise ratio can only be computed from consecutive errors. Nevertheless, much data is thrown away by only computing the first K_m errors. It is well possible that computing α over the another set of errors gives a totally different value of α.

3.3. Another Problem Regarding the Reliability of Average ERPs

From the OH-method a number of errors needed to have internal consistency (k^*) is determined. Therefore, it is known that exactly the first k^* error trials give internal consistent ERPs, such that average ERPs should be computed as average over the first k^* brain activity values. However, in practice, often the average ERPs are decided by averaging all error trials of a participant. Averaging over all trials only gives reliable average ERPs if indeed adding an error trial to the set of k^* errors does not significantly change the average ERP. This means that ERN and Pe values for the error trials of a participant should be independent over the error trials. There are two issues that determine this independency.

First, suppose that brain activity is higher when making the first error than when making the last error, because one gets accustomed to making errors. Then, ERPs are dependent over time, such that the average based on the first k^* errors can be substantially different from the average based on all errors. If this is the case, averages are not reliable as adding more trials can result in more errors and thus, in lower averages.

Secondly, suppose that different stimuli can give different brain activity values. Reliability is decided on the stimuli included in the first k^* errors. Including

all errors will probably change the share of different stimuli. Therefore, reliability of all errors can differ from reliability of the first k^* errors. In the same way, average ERPs can differ.

These issues show that it needs to be justified that the average includes all error trials. If this would be justified, the present procedure in the literature is correct. That is, first, the number of errors (k^*) needed for internal consistent ERPs is found. Then, this number of errors is used as a selection criterion to exclude all participants with less than k^* errors. After that, the average ERPs over all error trials are computed and finally, these averages are related to psychological disorders. However, if averaging all error trials do not give reliable average ERPs, finding a selection criterion is useless and thus additional research on the number of errors to include in the average should be done.

3.4. Research Goals

The OH-method is used to compute Cronbach's α. From this α the number of errors needed for internal consistency, k^*, is concluded. For example, if the α based on $k = 4$ errors, that is, computed using the first four columns of X, is high enough following Table 2, we need $k^* = 4$ errors for each person to get reliable results. However, as Cronbach's α is computed in the wrong way, this number of errors could possibly be biased. The main purpose of this research is to investigate whether the two problems mentioned in Section 3.2 bias the number of errors (k^*) needed for internal consistency.

Furthermore, k^* will be used to decide upon which participants to include in the ultimate sample. Then, averages over all errors are considered. However, averaging all errors does not necessarily give reliable average ERPs. Therefore, another research goal is to justify that taking the average of all error trials give reliable estimates.

This justification will be examined first. Then, if it is justified to average over all error trials, the average ERPs are reliable, and thus, it becomes useful to investigate the selection procedure (the main goal).

3.5. Approaches to Investigate Research Goals

This section gives an overview of approaches that we will use to investigate the research goals. We will start focusing on whether it is justified to average over all error trials. This justification is done using an empirical data set, which will be considered first. Then, we will examine the two issues discussed in Section 3.3. For the time dependency issue, we will test whether ERPs contain significant trends over time. For the other issue, we will use a random parameter model to test whether different stimuli give different brain activity values.

If averaging over all error trials is justified, the selection procedure can be considered. Previous section showed the two problems arising from the OH-method. Namely, assumptions $\mathcal{A}.1$ and $\mathcal{A}.2$ are violated and only one specific set of error trials is considered. To examine whether these problems bias the number of errors coming from the OH-method, we consider some approaches.

First, the empirical data set is used to obtain empirical distributions for Cronbach's α. Using these distributions, we can determine whether taking one

specific set of errors gives a reliable estimate of α. However, empirical data, which looks like the matrices of Tables 3 and 4, cannot fulfil the assumptions $\mathcal{A}.1$, $\mathcal{A}.2$, and $\mathcal{A}.3$ simultaneously. Therefore, different combinations of $\mathcal{A}.1$, $\mathcal{A}.2$, and $\mathcal{A}.3$ are assumed to obtain the empirical distributions.

Further, to investigate whether violating $\mathcal{A}.1$ and $\mathcal{A}.2$ gives bias, 'complete' data, that fulfils $\mathcal{A}.1$, $\mathcal{A}.2$, and $\mathcal{A}.3$, is needed. Using this 'complete' data set, a 'true' Cronbach's α, say α_T, can be computed. Then, α_T can be compared to α coming from the OH-method, say α_{OH}, to examine whether α_{OH} is a reliable estimate of α_T.

A 'complete' $N \times K$ data matrix Y should meet assumptions $\mathcal{A}.1$, $\mathcal{A}.2$, and $\mathcal{A}.3$. To fulfil these assumptions, two issues need to be ensured. First, matrix Y should not contain missings. A missing in the empirical data corresponds to a correct trial. To avoid these missings, participants need to fail all trials. Additionally, participants should face the same randomisations over the trials, such that each participant faces the same order of stimuli. Having this matrix Y ensures that Cronbach's α can be computed without doing any transformations at beforehand, such that the problems of the OH-method do not occur. However, we cannot ensure participants to fail all trials. Besides, most empirical data sets assign the type of stimulus randomly for each participant. Hence, empirical data usually lacks both essential requirements for computing α_T.

It will be impossible to ever get data matrix Y in practice, as there will always be participants that respond correct on some trials. Therefore, a simulation study will be explored. The simulation can generate a matrix Y, where respondents fail in each trial and where randomisations will be the same for everyone. From such simulated data, α_T can be computed. Then, an accordingly constructed α_{OH} can be calculated and α_T can be compared with α_{OH} to investigate possible bias.

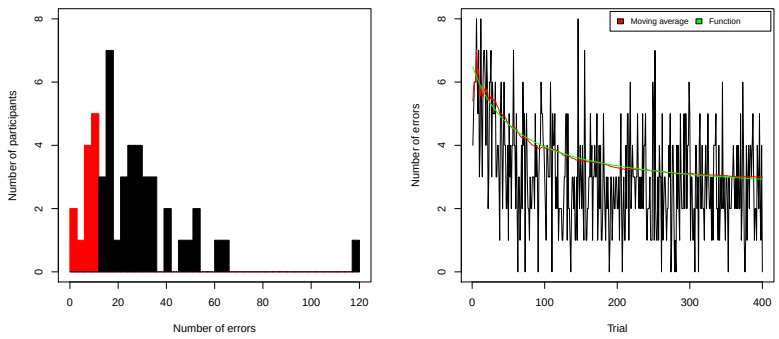
4. Data

In this section, we will introduce the empirical data set that will be used for justification and for the empirical distributions. We will first show the particular experiment set up used for this data. Then, we give some insightful details of the data.

The data consists of $N = 51$ healthy students of the Erasmus University Rotterdam, collected in September 2013 by Wim Rietdijk, PhD student at Erasmus University, and prepared by myself. The graduate students participated in an electroencephalographic (EEG) experiment in the Erasmus Behavioural Lab. To be specific, they participated in an Eriksen Flanker experiment. Each student considers 408 Eriksen Flanker tasks. From these tasks, the first 8 are burn-in trials and thus excluded from the analysis. This leaves $K = 400$ trials per participant, each taking no more than two seconds. During a trial one of the four stimuli as explained in Section 2.3 is shown and participants need to respond by pressing a '1' or a '5'. Brain activity is recorded for 32 electrodes placed along the scalp. From those recordings, ERN and Pe values are created for each student in each error trial. However, as these measures are associated

with errors, we are only interested in wrong responses. Therefore, values for correct responses are considered as missing. In terms of the data, this means that they are set equal to zero. This procedure gives two $N \times K$ matrices, one for ERN and one for Pe values, each containing many zeros and some brain activity values for those trials that resulted in an error. The sparsity of these matrices is 8.14 percent, meaning that the matrices contain many zeros. To visualize the matrices, consider the example matrices in Tables 3 and 4.

In order to reconstruct the OH-method, the number of errors made by a participant should at least be equal to $K_m = 14$. Because some participants made fewer than K_m errors, we deleted those from our data resulting in $N = 37$ participants. As each participant faces different randomisations of the trials, there is never a participant that considers exactly the same order of stimuli as one another. Nevertheless, every participant has 100 trials of each of the four possible stimuli from Table 1. Also, it happens that students neither press '1' nor '5'. For those observations, the resulting ERN or Pe values will be treated as correct responses, such that they are set equal to zero. Further, the Signal-to-Noise ratios for electrode averages for all trials of a participant for FCz and Pz are 0.30 and 0.28, respectively. This is in line with previously reported studies. For details, review Olvet and Hajcak (2009), Meyer et al. (2013), Rietdijk et al. (2014). Finally, in the empirical data, we have N individuals, R different stimuli, and K measurements of each individual. Therefore, we can consider the empirical data as panel data.



(a) The number of errors made by included participants (black) and excluded ones (red).

(b) The number of errors made for each trial together with a moving average and function.

Figure 4: Total number of errors.

Now, we will consider some characteristics of the data. Figure 4 shows a histogram of the number of errors for the participants and a graph of the number of errors over time. The histogram in Figure 4a shows the number of errors made by included participants (black) and the excluded ones (red). There is one extreme outlier, namely, a student with 119 errors. Furthermore,

the minimum number of errors made in the black part of the histogram is equal to $K_m = 14$, as, of course, everyone with less errors is excluded from the data. Also, the median of the errors for the included participants is equal to 28. Figure 4b shows the number of errors made for each trial, or, the number of errors as a function of time. From this figure, it can be inferred that, although participants had eight burn-in trials, they still make more errors in the beginning of the experiment than at the end of the experiment. The red line in Figure 4b shows a moving average with a window of 11 trials to provide a smoother trend. This line shows that the number of errors decreases over time and eventually stabilizes. The green line is defined by $y = \frac{a}{x+b} + c$ with $a = 253.20$, $b = 60.69$ and $c = 2.38$. The parameters of this function are decided by minimizing the sum of the squared differences between the moving average and the function. The function models the moving average quite well. We will use this function to model the number of errors over the trials in the simulation study later on.

Table 5: The total number of errors and their proportions corresponding to the different stimuli.

Stimuli	Number of errors	Proportion(%)
SSSSS	138	11.5
HHSHH	434	36.0
HHHHH	145	12.0
SSHSS	488	40.5

Table 5 shows the total number of errors made in each stimulus, together with their proportions. There is a clear difference in the number of errors made in incongruent and congruent stimuli. The incongruent stimuli contain many more errors than the congruent ones. However, there are minor differences within the two congruent stimuli and the two incongruent stimuli.

Figure 5 shows the error pattern for each participant. In this figure, each value on the y-axis represents one student. The dots corresponding to each student, that is, between two black lines, show error trials. Note that there is a distinction between the four stimuli. This figure suggests that errors come in groups. Therefore, considering each participant separately, the probability of making an error is dependent of the current past.

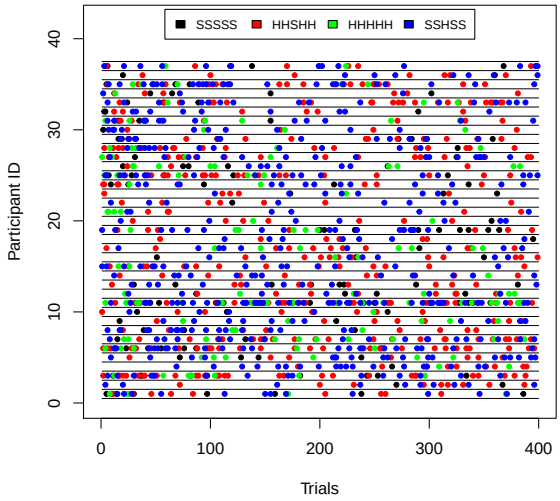


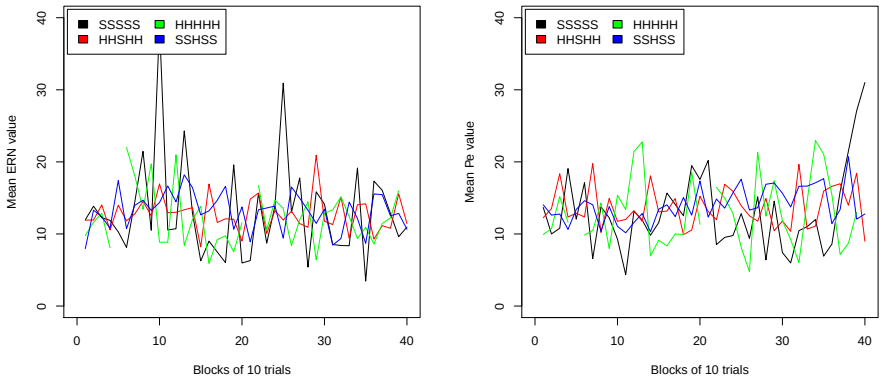
Figure 5: The error trials for each participant.

5. Justification of Independency of ERPs Within Participants

In this section, justification of averaging over all trials will be considered. For averaging all trials to be justified, ERN and Pe values should be independent over the error trials of a participant. There are two issues that constitute this independency. Namely, the ERPs should not be time dependent and different stimuli should not give different ERPs. We start considering whether ERPs are time dependent. Then, we set up a random parameter model to test whether different stimuli result in different ERPs.

5.1. Determining Time Dependency of ERPs

In this section, we will consider whether ERPs are time dependent, where time dependency means that ERPs in the beginning of the experiment differ from ERPs in the end of the experiment. To do so, consider Figure 6, where the average ERN and Pe values are plotted against 40 blocks of 10 trials. The figures show that there is not much difference between ERN or Pe values. Furthermore, note that some values are missing because no errors of a specific stimuli happened in a block. This causes some lines in the figures to be interrupted.



(a) ERN (b) Pe

Figure 6: The average value for ERN or Pe over 40 blocks of 10 trials.

To find whether there is a trend in the averages in Figures 6a and 6b, we will explore a regression analysis. The regression analysis is explained for one stimulus for ERN values. That is, it is explained for one line in Figure 6a. Define the average ERN value for one stimulus per 10 trials as $\bar{x}_b$, $b = 1, ..., 40$, where b denotes the particular block over which the average is computed, and, thus, indicates time. If a particular block do not contain errors of the stimulus that is considered, an average cannot be computed, resulting in a missing value. This missing value for block b, will be estimated as the average of block $b-1$ and $b+1$ ($\bar{x}_b = \frac{\bar{x}_{b-1} + \bar{x}_{b+1}}{2}$). Then, $\bar{x}_b$ is regressed on a constant and a trend (b). This gives the following model

$$\bar{x}_b = \tau_0 + \tau_1 b + \varepsilon, \tag{2}$$

where the error term is assumed have a normal distribution. Now, preforming a regression analysis results in p-values of the trend parameter (τ_1), which are shown in Table 6. Most results are insignificant based on a five percent significance level, except for the average Pe value for 'SSHSS'. However, there is an outlier in the 38^{th} block, which can cause this significant result.

Table 6: The p-values for the significance of the trend parameter (τ_1) in Equation (2).

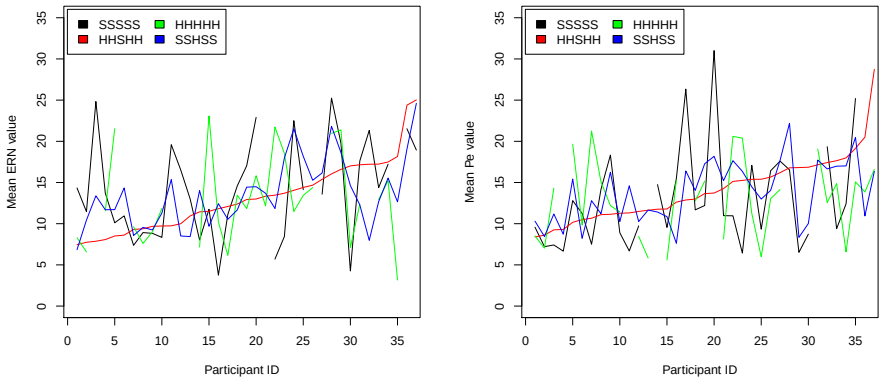
Stimuli	ERN	Pe
SSSSS	0.57	0.18
HHSHH	0.71	0.83
HHHHH	0.47	0.64
SSHSS	0.36	0.00

As most trend parameters are insignificant, average ERN and Pe values do,

on average, not change over time. This means that an average of the ERN or Pe is not influenced by time, or, the average ERPs are not time dependent.

5.2. Testing Differences in Brain Activity Across Different Stimuli

As Section 4 showed that the number of errors for incongruent and congruent stimuli differ substantially, it can be asked whether this is also visible in the corresponding ERN and Pe values. So, we will examine whether different stimuli do give different ERPs. Therefore, we first consider Figure 7a (7b), which shows the average ERN (Pe) value per participant for the different stimuli. Again, figures for ERN and Pe do not differ that much. Also, lines can be interrupted because some participants do not fail a specific stimulus.



(a) ERN (b) Pe

Figure 7: The average ERN or Pe value for each participant sorted on 'HHSHH'.

The ERN and Pe values in those figures are sorted on 'HHSHH', such that a slightly upward trends becomes visible in the stimuli other than 'HHSHH'. To test whether there is association in the four averages in the figures, Spearman's correlation is used. The correlations and corresponding p-values are shown in Table 7. Many correlations are significant based on a five percent significance level. Especially incongruent trials show significant results. The significant results mean that a participant with high ERN or Pe values on one stimulus also has a higher ERN or Pe value on the other stimulus. Therefore, ERPs are individual specific.

Table 7: The correlation matrix (lower triangle) with corresponding p-values (upper triangle) for ERN and Pe.

	ERN				Pe			
Stimuli	1	2	3	4	1	2	3	4
1 (SSSSS)	-	0.04	0.09	0.03	-	0.06	0.42	0.00
2 (HHSHH)	0.36	-	0.09	0.00	0.32	-	0.52	0.00
3 (HHHHH)	0.34	0.31	-	0.02	0.16	0.12	-	0.02
4 (SSHSS)	0.38	0.58	0.42	-	0.54	0.51	0.42	-

To test whether different stimuli give different brain activity values, we will set up a model for brain activity. So, the outcome of this model should be an ERN or Pe value. This ERN or Pe value should be declared with the available information, that is, the type of stimulus that is considered in a particular trial. As Figure 7 and Table 7 showed, there are significant trends in ERN and Pe values over the participants. This means that if a particular respondent has a high ERN or Pe value on one stimulus, he or she will also have higher ERN or Pe values on other stimuli. Therefore, a model with individual-specific coefficients is appropriate. Let x_{ij} be defined as in Section 3.1 and consider the random parameter model

$$\log(x_{ij}) = d_j'\beta_i + \varepsilon_{ij}, \tag{3}$$

where d_j is a $R \times 1$ vector, with R the number of different stimuli. Note that we take $d_{j1} = 1 \ \forall \ j$, so that the first term in $d_j'\beta_i$ is an intercept. The elements d_{jr} for $r = 2, 3, 4$ are defined as

$$d_{jr} = \begin{cases} 1 & \text{if trial } j \text{ is of stimulus type } r, \\ 0 & \text{otherwise.} \end{cases}$$

We need to let out one stimulus to prevent multicollinearity and we choose to let 'SSSSS' out as this stimulus contains the fewest errors (see Table 5). Furthermore, we assume $\beta_i \sim NID(\beta, \Sigma_\beta)$ with Σ_β an $R \times R$ covariance matrix and $\varepsilon_{ij} \sim NID(0, \sigma_\varepsilon^2)$.

To be able to test whether different stimuli give different brain activity values, we will estimate the parameter of Equation (3) using the empirical data of Section 4. To estimate the parameters, rewrite Equation (3) as

$$\begin{aligned} \log(x_{ij}) &= d_j'\beta + d_j'(\beta_i - \beta) + \varepsilon_{ij} \\ &= d_j'\beta + u_{ij}, \end{aligned}$$

with $u_{ij} = d_j'(\beta_i - \beta) + \varepsilon_{ij}$, such that $E[u_{ij}] = 0 \ \forall \ i, j$ and $E[u_{ij}u_{is}] = d_j'\Sigma_\beta d_s + \sigma_\varepsilon^2 I[j = s]$. In this setting, β can be estimated with a Feasible Generalized Least-Squares (FGLS) estimator, see Swamy (1970). For this purpose, we write the model in matrix notation as

$$\log(X_i) = D_i\beta + u_i.$$

Here, X_i is a $K_i \times 1$ vector with brain activity, where K_i is the number of error trials for participant i, D_i is a $K_i \times R$ matrix with the values d_j in it, and $u_i \sim N(0, \Omega_{u_i})$ with $\Omega_{u_i} = D_i \Sigma_\beta D_i' + \sigma_\varepsilon^2 I$.

The FGLS estimator is given by

$$\hat{\beta} = \left(\sum_{i=1}^{N} D_i' \hat{\Omega}_{u_i}^{-1} D_i \right)^{-1} \left(\sum_{i=1}^{N} D_i' \hat{\Omega}_{u_i}^{-1} \log(X_i) \right),$$

for which an estimate of Ω_{u_i} is necessary. To obtain $\hat{\Omega}_{u_i}$, we use the fact that $u_i \sim N(0, \Omega_{u_i})$ together with a concentrated log likelihood function. That is, we only need to get estimates for Σ_β and σ_ε^2. For this purpose, starting values for Σ_β and σ_ε^2 are taken, such that $\hat{\Omega}_{u_i}$ can be obtained. Then, we can find $\hat{\beta}$. Now, we can estimate Σ_β and σ_ε^2 using the following concentrated log likelihood function (apart from a constant) where we use $\log(X_i) \sim N(D_i\beta, \Omega_{u_i})$

$$\begin{aligned}
\log(f(\log(X)|\Sigma_\beta, \sigma_\varepsilon^2)) &= \log \left(\prod_{i=1}^{N} f(\log(X_i)|D_i, \hat{\beta}, \Sigma_\beta, \sigma_\varepsilon^2) \right) \\
&= -\frac{1}{2} \sum_{i=1}^{N} \log \left(|\Omega_{u_i}| \right) \\
&\quad - \frac{1}{2} \sum_{i=1}^{N} (\log(X_i) - D_i\hat{\beta})' \Omega_{u_i}^{-1} (\log(X_i) - D_i\hat{\beta}).
\end{aligned}$$

Maximizing this log likelihood function gives again estimates for Σ_β and σ_ε^2, such that the procedure can be repeated. We iterate until the estimates of β converge.

Applying this procedure to the empirical data of Section 4 gives the parameter estimates as below. From these estimates, it can be inferred that there is some mean value for the logarithm of ERN ($\hat{\beta}_1 = 2.44$) and for the logarithm of Pe ($\hat{\beta}_1 = 2.37$) and that switching stimuli will probably not influence the this mean value as the estimates for β_2, β_3, and β_4 are close to zero. Besides, $\hat{\sigma}_\varepsilon^2 = 0.40$ for ERN and $\hat{\sigma}_\varepsilon^2 = 0.27$ for Pe.

$$\hat{\beta}^{ERN} = \begin{bmatrix} 2.44 \\ 0.10 \\ 0.09 \\ -0.10 \end{bmatrix}, \hat{\Sigma}_\beta^{ERN} = \begin{bmatrix} 1.04 & 0.19 & -0.17 & 1.55 \\ 0.03 & 0.20 & -0.12 & 0.54 \\ -0.36 & 0.14 & 0.44 & 0.18 \\ 0.21 & 0.17 & 0.09 & 0.75 \end{bmatrix}$$

$$\hat{\beta}^{Pe} = \begin{bmatrix} 2.37 \\ 0.08 \\ -0.14 \\ 0.13 \end{bmatrix}, \hat{\Sigma}_\beta^{Pe} = \begin{bmatrix} 0.74 & 0.36 & -0.26 & 0.31 \\ -0.21 & 1.23 & 0.09 & 0.44 \\ -0.63 & 0.47 & 0.55 & 0.43 \\ 0.46 & 0.22 & 0.28 & 0.71 \end{bmatrix}$$

To test whether indeed the intercept for the logarithm of ERN or Pe differs significantly from zero and to test whether different stimuli give different brain

activity values, we use the Hoteling's T^2 statistic (Johnson et al., 1992, pp. 210-216). Therefore, we first use the fact that transformations $A\beta$ are also normally distributed with mean $A\hat{\beta}$ and variance $A\hat{\Sigma}_\beta A'$. For example, $\beta_1 \sim N(\hat{\beta}_1, \hat{\Sigma}_\beta[1,1])$, where $\hat{\Sigma}_\beta[1,1]$ is the element of $\hat{\Sigma}_\beta$ in the first column and in the first row. In the same way, using $A = [\mathbf{0}_3 \vdots I_3]$, we obtain $[\beta_2, \beta_3, \beta_4]' \sim N([\hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4]', \hat{\Sigma}_\beta[-1,-1])$, where $\hat{\Sigma}_\beta[-1,-1]$ is the block matrix of $\hat{\Sigma}_\beta$ in which the first row and the first column are deleted.

Now, testing significance for the intercept (β_1) boils down to the univariate case, where we can just use the t-test. The t-statistic for testing the null hypothesis of $\beta_1 = 0$ is equal to 5.92 for ERN and 6.80 for Pe. Using a five percent significance level, the critical value is 2.02 meaning that both null hypotheses can be rejected. Therefore, on average, the logarithm of the mean ERN (Pe) value differs significantly from zero.

Next, we will test the null hypothesis of $[\beta_2, \beta_3, \beta_4]' = \mathbf{0}_3$ using the Hotelling's T^2 statistic. It turns out that the statistic for ERN equals 0.23 and the statistic for Pe 0.13. The corresponding critical value, coming from the F-distribution and with a five percent significance level equals $\frac{(N-1)p}{N-p} F_{p,n-p} = 9.02$, where $p = 3$ is the number of parameters to test. From this we conclude that both null hypotheses are not rejected. Therefore, the impact of different stimuli on the brain activity value is not significantly different from zero, meaning that it is plausible that different stimuli do not generate different brain activity values. Therefore, ERPs are independent of the type of stimulus. Consequently, averaging ERPs over all error trials gives reliable averages.

5.3. Conclusions

From this section it can be concluded that ERPs are not time dependent. Also, different stimuli will not constitute to different brain activity values. Therefore, it is plausible that the ERN and Pe values within a participant are independent over the error trials of this participant. Consequently, considering the average over all error trials, instead of a selection of error trials, is justified. As a result, it is necessary to decide upon the number of errors needed for internal consistency, such that a selection criterion can be used to decide for which participants reliable averages can be obtained.

6. Empirical Distributions

In this section, we will consider ways to use the empirical data of Section 4 for computing Cronbach's α. As it will be impossible to construct a 'true' α from this data, we will compute different variants of α. We start by computing α_{OH}. Then, the focus will be on taking a random set of errors (instead of the first K_m errors) for computing α. However, considering a single random draw of errors will not give information about possible bias. Therefore, we would like to consider all possible random draws such that we get an exact empirical distribution of α. However, as obtaining α for all permutations is

computationally prohibitive, only $W = 10000$ random draws are performed. The empirical distributions, coming from these draws, will be based on different combinations of the assumptions $\mathcal{A}.1$, $\mathcal{A}.2$, and $\mathcal{A}.3$ of Section 3.2.

6.1. Methods

In this section, the methods will be discussed, starting with the OH-method. Then, we will explain how we choose a random draw of errors to obtain different values of Cronbach's α. Finally, we consider different methods to obtain empirical distributions.

6.1.1. Cronbach's α for the OH-method

We start with computing Cronbach's α using the OH-method that was explained in Section 3.1. In short, the first k error trials of each participant are taken, where $k \in \{2, 4, ..., K_m = 14\}$. This is possible as participants with less than 14 errors were excluded from the empirical data. Then, we compute α_{OH} for each k and for both ERN and Pe.

6.1.2. A Single Random Draw of Error Trials

In the OH-method, the first k errors are used to compute Cronbach's α. However, as the median number of errors made by respondents is equal to 28, many error trials are ignored. Therefore, we will consider α_R, based on a random draw out of the errors for each participant. In this way, each error of a participant can be used with equal probability.

We will explain the procedure for getting a random draw very carefully as it will be important for explaining the methods for creating empirical distributions later on. In the OH-method, α_{OH} is obtained for each k and for both ERN and Pe. This gives a total of 14 values of α. Therefore, considering a single random draw of error trials should also give 14 values of α, or, a random value for α is needed for each value of k and for both ERN and Pe. Comparison over these different values for k and over ERN and Pe values is desirable because of the following reasons. First, having comparable values for different values of k makes it possible to draw conclusions over the number of errors needed for an α that is high enough. Secondly, ERN and Pe values are jointly recorded, so that they naturally belong to each other such that α for ERN and Pe values should be based on the same error trials.

To create comparable values for α, there has to be one specific $N \times K_m$ matrix with random error trials for each participant. From this matrix, matrices with ERN and Pe values can be found by translating the random error trials to the brain activity values. Then, considering the matrix for ERN (Pe), the first $k = 2$ columns, or, error trials, are taken to compute α_R for $k = 2$. After that, the first $k = 4$ errors of the same matrix are taken to compute α_R for $k = 4$. This is done for all $k \in \{2, 4, ..., K_m\}$ and for both the ERN and Pe matrix.

To get the $N \times K_m$ matrix with random error trials for each participant, the following is done. All error trials for one participant are taken. From those error trials, we take a random draw of size K_m without replacement. Doing

this for each participant results in a $N \times K_m$ matrix with random error trials. Note that the random draw is not sorted, such that errors in this matrix are not necessarily chronological.

Note that α_R is not valid in the sense that assumption $\mathcal{A}.1$ and $\mathcal{A}.2$ of Cronbach's α are still violated. In contrast to the OH-method, assumption $\mathcal{A}.3$ is also violated here as the error numbers do not necessary have to correspond to each other. By creating the empirical distributions, which are all based on W random draws of errors, or, on W values of α_R, we will fulfil different combinations of $\mathcal{A}.1$, $\mathcal{A}.2$, and $\mathcal{A}.3$.

6.1.3. Empirical Distributions: Method A

The first way to create a distribution for Cronbach's α is based on trying to be as liberal as possible. Namely, we assume that Cronbach's α is robust for violation of assumptions $\mathcal{A}.1$, $\mathcal{A}.2$, and $\mathcal{A}.3$. Actually, Method A violates even more assumptions than the OH-method, where at least $\mathcal{A}.3$ is fulfilled. To make these distributions possible, we perform repeatedly the random draw procedure as explained in Section 6.1.2.

6.1.4. Empirical Distributions: Method B

To come one step closer to the OH-method, while still trying to be as liberal as possible, consider method A again. Method B is the same as Method A except for that only the first K_m error trials (instead of all error trials) are permuted.

6.1.5. Empirical Distributions: Method C

Methods A and B assume that Cronbach's α is robust for violation of assumptions $\mathcal{A}.1$, $\mathcal{A}.2$, and $\mathcal{A}.3$. However, in the OH-method, the assumption $\mathcal{A}.3$ is fulfilled. Therefore, Method C will also try to fulfil this assumption. Remember from Section 3.2 that assumption $\mathcal{A}.3$ is fulfilled as the first column of the transformed matrix contains errors that were made first by the participants, the second columns of the transformed matrix contains errors that were made secondly by the participants, and so on. To meet this assumption, there is only one possibility for the selection of errors, namely the errors as used in the OH-method. Therefore, to get multiple draws instead of just one, we cannot exactly meet assumption $\mathcal{A}.3$, but we try to approximate it. To do so, the procedure for getting a random draw as described in Section 6.1.2 is slightly changed. Namely, the order of the errors is maintained while selecting K_m errors out of all errors for each participant, such that there is at least a natural ordering of the errors in computing α.

6.1.6. Empirical Distributions: Method D

Another way to create an empirical distribution for Cronbach's α includes controlling for the type of stimuli ($\mathcal{A}.1$) and taking into account the order of the errors ($\mathcal{A}.3$). Method C shows that we only can approximate assumption $\mathcal{A}.3$. In Method D, assumption $\mathcal{A}.1$ is fully met and $\mathcal{A}.3$ partially, such that this

Method satisfies most assumptions of Section 3.2. To meet assumption $\mathcal{A}.1$, we will only distinguish between congruent and incongruent stimuli to avoid losing too many observations. Here, the procedure of Section 6.1.2 is again slightly changed. To get a random draw of error trials for a participant, two random draws are taken, that is, a random draw of $\frac{K_m}{2}$ congruent and a random draw of $\frac{K_m}{2}$ incongruent error trials. Both random draws are sorted to obtain the natural order ($\mathcal{A}.3$). Then, we create the $N \times K_m$ random draw matrix, where we have congruent trials in the first column, incongruent trials in the second one, again congruent trials in the third column, and so on. The reason for alternating congruent and incongruent stimuli is that, in this way, computing Cronbach's α based on $k < K_m$ will include as many congruent as incongruent stimuli. This makes comparison over the number of errors (k) more fair. Note that, because every participant needs to have at least $\frac{K_m}{2}$ congruent and $\frac{K_m}{2}$ incongruent errors, the number of usable students decreases to $N_m = 17$, such that our random matrix is of size $N_m \times K_m$.

6.2. Results

In this section, we consider the results, starting with α_{OH}. We do not give results of α_R as having α based on just one random draw of errors will not give insightful results. Then, the empirical distributions are discussed. All figures containing histograms, also contain the corresponding value of α_{OH}, the median of the histogram, α_M, and the confidence interval of the histogram. These empirical confidence intervals are decided by taking 95 percent of the middle values of the histograms.

6.2.1. Cronbach's α for the OH-method

Table 8 shows the values for Cronbach's α using the OH-method for ERN and Pe and based on different values for the number of errors (k). The table demonstrates that α_{OH} mostly increases with the number of errors. Also, α_{OH} for ERN is substantially larger than α_{OH} for Pe. Olvet and Hajcak (2009) find a moderate α (see Table 2) for ERN with $k = 6$ and for Pe with $k = 2$ with a total of 53 participants. Considering the results of the empirical data for ERN in Table 8, we find a moderate α of 0.58 with $k = 8$, but for Pe, even with $k = 14$ we do not have enough error trials to have at least a moderate α ($\alpha > 0.50$). However, the empirical data used here has fewer participants than the data used in Olvet and Hajcak (2009), which can cause these differences.

Table 8: Cronbach's α using the OH-method, α_{OH}, for the empirical data for ERN and Pe based on a different number of errors (k).

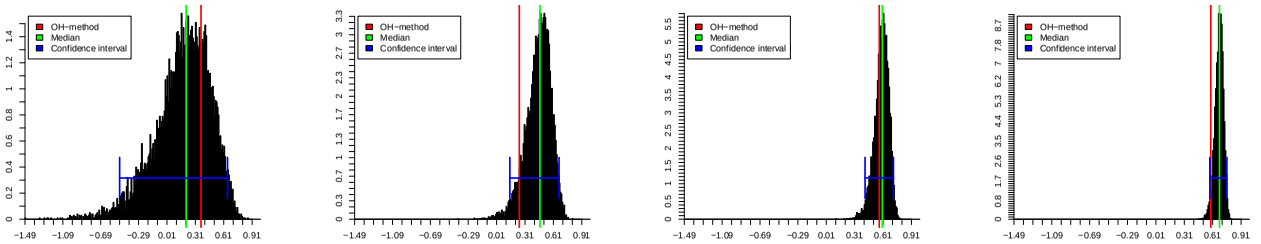
	α_{OH}	
k	ERN	Pe
2	0.37	0.02
4	0.15	-0.17
6	0.25	0.12
8	0.58	0.31
10	0.57	0.28
12	0.58	0.38
14	0.59	0.45

Furthermore, in practice, it is often recommended that α should at least be higher than 0.70. However, considering, for example, ERN values and taking 0.70 as a threshold, gives 26 errors. Note that, to obtain this result, K_m should be equal to 26, such that all participants with less than 26 errors need to be excluded from analysis. This leaves $N = 22$ participants, such that many participants are ignored.

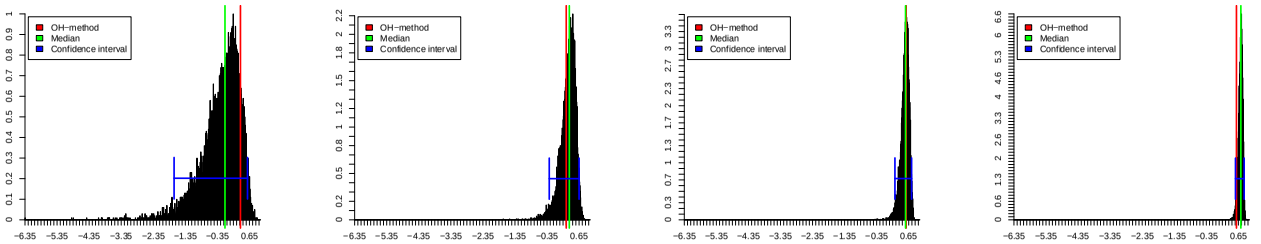
6.2.2. Empirical Distributions: Method A

In Figure 8, we see the histograms coming from Method A, where we tried to be as liberal as possible, corresponding to ERN in the upper row for the values $k \in \{2, 6, 10, 14\}$. We see that with increasing the number of errors, k, the histograms become smaller. This shrinkage is caused by an increasing lower bound. This is not surprisingly as α cannot exceed one. However, the histograms are very wide, also with higher values of k, indicating that different random draws of errors can differ the value of α a lot. Comparable results are found for Pe.

Also, α_{OH} is, in most cases, substantially lower than α_M. This effect indicates that α_{OH} mostly underestimates Cronbach's α in Method A, which can also be seen in Table 9. Note that a positive difference $(\alpha_M - \alpha_{OH})$ indicates an underestimation of α_M. The differences for ERN and Pe are large in most cases, so that α_{OH} clearly underestimates the median of the distribution. Also, the upper bound of the histograms does not increase that much if k increases. Nevertheless, the lower bound does increase, but we would still have completely different conclusions in lower bound or upper bound cases. For instance, if we would like to have a moderate value of α $(0.5 < \alpha \leq 0.7)$ and if we consider the results for Pe, we conclude that, considering the lower bound, 14 errors are still not enough for a moderate α (as $\alpha = 0.45 \leq 0.50$) while, considering the upper bound, having $k = 2$ errors will already be enough for a moderate α (as $\alpha = 0.59 \geq 0.50$).



(a) Method A, ERN, $k = 2$ (b) Method A, ERN, $k = 6$ (c) Method A, ERN, $k = 10$ (d) Method A, ERN, $k = 14$



(e) Method D, ERN, $k = 2$ (f) Method D, ERN, $k = 6$ (g) Method D, ERN, $k = 10$ (h) Method D, ERN, $k = 14$

Figure 8: The distribution of Cronbach's α for the ERN value for Method A and D and for different values of k, together with α_{OH}, α_M, and the corresponding confidence intervals.

Table 9: The differences $\alpha_M - \alpha_{OH}$ for different values of k, together with the lower bound (L-bound) and upper bound (U-bound) of the confidence intervals for Method A.

	ERN			Pe		
k	$\alpha_M - \alpha_{OH}$	L-bound	U-bound	$\alpha_M - \alpha_{OH}$	L-bound	U-bound
2	-0.16	-0.49	0.65	0.13	-0.52	0.59
4	0.22	-0.08	0.63	0.45	-0.17	0.56
6	0.22	0.15	0.67	0.25	0.04	0.60
8	-0.03	0.30	0.70	0.14	0.18	0.62
10	0.03	0.42	0.72	0.22	0.29	0.65
12	0.07	0.51	0.74	0.17	0.38	0.67
14	0.09	0.58	0.76	0.14	0.45	0.69

So, assuming that Cronbach's α is robust for the violation of assumptions $\mathcal{A}.1$, $\mathcal{A}.2$, and $\mathcal{A}.3$, there is an underestimation of the median of the empirical distributions relative to the OH-method. Also, histograms are still wide with 14 errors such that we would have different conclusions about the number of errors needed for internal consistency in the lower bound case and in the upper bound case.

6.2.3. Empirical Distributions: Method B

In Method B, the first K_m errors were permuted W times. The conclusions for $k < K_m$ are roughly the same as in Method A. However, for $k = K_m = 14$ errors, we get a very small histogram for ERN (see Figure 9) as the confidence interval is $[0.58, 0.60]$. The same result holds for Pe.

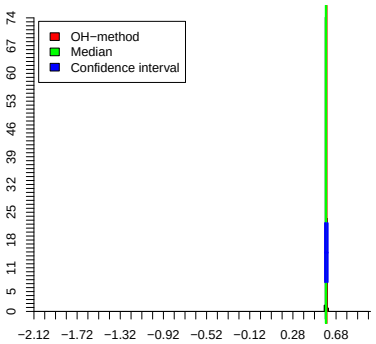


Figure 9: The distribution of Cronbach's α for the ERN value for Method B and $k = K_m$, together with α_{OH}, α_M, and the corresponding confidence intervals.

This fact means that permuting the first K_m errors and taking k equal to

K_m will not differ α substantially. However, taking k smaller than K_m does. To be clear, in the special case of $k = K_m$ (thus permuting exactly the first 14 errors of the empirical data), the values in the matrix for computing α are exactly the same in each random draw, but they are placed in a different order. In this case, the sum scores are the same in each repetition. Also, the sum of the variances of the columns does not differ that much (range = [980, 1040]), such that, considering one participant, the magnitude of ERPs will probably be roughly the same over the errors. This is in line with the fact that ERPs are independent of the error trials as concluded in Section 5.3. The results of Method B suggest that violation of the assumptions does not influence Cronbach's α, but the particular set of errors that is included does.

6.2.4. Empirical Distributions: Method C

The histograms for Method C are comparable with those of Method A (where we tried to be as liberal as possible), such that we roughly get the same conclusions. In Table 10, we see the differences between α_M and α_{OH} together with the lower and upper bound of our confidence intervals. In comparison with Method A, the confidence intervals for this method are in a lower range for lower values of k, but when k increases, the intervals become more and more the same. Therefore, intervals are again substantially such that having another set of errors for computing α can give totally different conclusions about the number of errors needed.

Table 10: The differences $\alpha_M - \alpha_{OH}$ for different values of k, together with the lower bound (L-bound) and upper bound (U-bound) of the confidence intervals for Method C.

	ERN			Pe		
k	$\alpha_M - \alpha_{OH}$	L-bound	U-bound	$\alpha_M - \alpha_{OH}$	L-bound	U-bound
2	0.01	-0.38	0.67	-0.10	-0.71	0.37
4	-0.01	-0.31	0.44	0.28	-0.37	0.43
6	0.08	-0.01	0.57	0.11	-0.12	0.47
8	-0.06	0.32	0.66	0.01	0.06	0.51
10	0.00	0.40	0.68	0.14	0.21	0.57
12	0.05	0.50	0.71	0.13	0.35	0.63
14	0.09	0.58	0.75	0.14	0.46	0.68

6.2.5. Empirical Distributions: Method D

In the lower part of Figure 8, the histograms for the most correctly created values of α are shown. Again, α_{OH} underestimates α_M in most cases. Further, if we compare Methods A and D, the lower bounds of Method D are much lower than those of Method A, while the upper bounds are roughly the same. This would indicate that α becomes less accurate if the assumption $\mathcal{A}.1$ and $\mathcal{A}.3$ are (partially) met. However, we need to note here that method D is based on fewer observations than the other methods, so that this can also be the reason for wider distributions.

6.3. Conclusions

Results in this section suggest that violation of the main assumption of Cronbach's α does not influence the value of α, but the specific set of error trials that is used for computation does.

First, when permuting the error trials in the rows of one specific random matrix, values for α will not differ substantially. Therefore, it can be suggested that the three assumptions mentioned in Section 3.2 are not important for getting a reliable α.

However, empirical distributions for Cronbach's α are very wide, such that there is much difference in lower bound and upper bound of the corresponding confidence intervals. Therefore, the specific set of errors taken to compute α, can influence the value of α, and thus, k^*, a lot. Namely, when considering the lower bound of the distributions as a reliable estimate of α, we would need many more errors (often more than 14 for Pe) to have a moderate α, than when considering the upper bound of the distributions, where a moderate α can often be concluded for $k = 2$.

7. Simulation

Section 3.1 discusses the violation of the sub assumptions $\mathcal{A}.1$ and $\mathcal{A}.2$ of Cronbach's α. This violation causes possible bias in the number of errors needed for internal consistency in the ERPs. Section 3.5 shows that one approach to investigate this bias, is comparing α_T to α_{OH}. However, it is impossible to compute α_T from empirical data, such that a simulation study is necessary.

The data generating process (DGP) of this simulation should contain two models. The first model should create the 'complete' matrix Y (see Section 3.5). This matrix permits that α_T can be computed. The second model should create errors to obtain the $N \times K$ matrix X (see Section 3.1). From this matrix X, one can compute an accordingly constructed α_{OH}. Comparing α_T and α_{OH} will enable us to find possible bias in the value of Cronbach's α computed with the OH-method, and thus, in the number of errors needed for internal consistent ERPs.

In the simulation study, different experimental conditions, that is, different ways to set the parameters of the DGP, can be considered. Simulating data for different experimental conditions enables to investigate whether the varying parameters significantly affect Cronbach's α.

Next to examining the bias in the number of errors needed for internal consistency in the ERPs and investigating the influential parameters, the simulation study can also be used to find other methods for establishing the reliability of ERPs. Therefore, we will consider another measure of internal consistency, namely, test-retest reliability and we propose a method based on the empirical distributions in Section 6.

In sum, the simulation study has three purposes. First, the possible bias coming from the OH-method is examined. Then, it is used to find the impact of several parameters on Cronbach's α, and finally, other methods for finding the number of errors needed for internal consistency are considered.

We will first define the DGP. The models used in this DGP are the same for ERN and Pe values. However, as parameter estimates from Section 5.2 differ between ERN and Pe, we allow parameters in these models to differ for ERN and Pe. Then, we specify experimental conditions. After that, we will consider outcomes of the simulation. Next, we consider the simulation study itself and we show simulation results. Finally, we draw conclusions from the simulation results.

7.1. Data Generating Process

To be able to compute both α_T and α_{OH}, the matrices Y and X are needed. In order to get those matrices, the data generating process should combine two models.

First, there needs to be a model to create 'complete' data Y. This model should generate brain activity values and meet the requirements to get the matrix Y, that is, participants should fail each trial and the same order of stimuli should be imposed on each student. This 'complete' matrix Y is obtained using the model for brain activity in Section 5.2. Running this model for all participants and all trials will lead to the matrix Y.

To create errors in the matrix Y, there needs to be a second model. The outcome of this model needs to be binary, as a trial can either be correct or it can be incorrect. So, to ultimately get the matrix X, a model that creates a binary matrix, say P^*, that indicates error trials with the value 1, is needed. Element wise multiplying this matrix P^* with Y will give X.

To achieve this, we will make a probability process which can be translated to this binary outcome. To get the probability process, we take several issues into account. First, students make more errors in incongruent trials than in congruent ones, see Table 5. As a result, a distinction in probability has to be made for different stimuli. However, the parameters for different stimuli do not have to be individual specific as we did for the model in Equation (3) because every student fails more incongruent trials than congruent ones. Then, as suggested by the histogram in Figure 4a, it should be allowed that participants differ in the number of errors that they make. Further, considering Figure 4b, we see that the total number of errors made slowly decreases over time. So, the probability of making an error in the beginning of the experiment should be somewhat higher than the probability of failing later on. Finally, as can be seen from Figure 5, errors occur mostly in groups. Therefore, the current history of the errors for a participant should be taken into account.

We first focus on a model for getting a probability of failing. That is, we consider one participant and try to find a probability of failing for each trial. Define the probability of failing trial j for participant i as p_{ij}. Then, we propose the following model

$$p_{ij} = \frac{\exp\left(\gamma z_j + d_j'\delta + \rho \sum_{l=j-4}^{j-1} p_{il} + \epsilon_{ij}\right)}{1 + \exp\left(\gamma z_j + d_j'\delta + \rho \sum_{l=j-4}^{j-1} p_{il} + \epsilon_{ij}\right)}, \tag{4}$$

where z_j is the function of Figure 4b to ensure more errors in the beginning of the experiment, d_j is a $R \times 1$ vector containing dummies for the stimuli, and $\sum_{l=j-4}^{j-1} p_{il}$ ensures that the current past is taken into account. Furthermore, we assume $\epsilon_{ij} \sim NID(0, \sigma_\epsilon^2)$. To simulate from this model, there is a sequential procedure over the trials because of $\sum_{l=j-4}^{j-1} p_{il}$. To start this procedure, there are four probabilities in the past needed. To get these probabilities, a random draw of a uniform distribution is taken for the probabilities of the first trial. Then, for the second, third and fourth probability, probabilities are generated with the available probabilities. So, to get the second probability, only the first probability is used and so on.

To get the final error process, it is required that the probability process just described is translated to a binary process. The black part of the histogram in Figure 4a is used to sample the number of errors that each participant makes. Note that we sample with replacement. This gives the number of errors ($\hat{K}_i$) for each simulated participant i. Note that $\hat{K}_i \geq K_m$. Each respondent has a $1 \times K$ vector p_i with probabilities from Equation (4). To decide upon the binary process, the $\hat{K}_i$ highest probabilities in p_i will be the errors in the binary process. To be precise, we define the binary outcome process p_{ij}^* as

$$p_{ij}^* = \begin{cases} 1 & \text{if } p_{ij} \text{ belongs to the highest } \hat{K}_i \text{ probabilities of } p_i, \\ 0 & \text{otherwise.} \end{cases}$$

From all values p_{ij}^* we obtain the binary $N \times K$ matrix P^*. Now, this matrix is multiplied element wise with Y to obtain X. To set the parameters of this model in the simulation study, we would like to have an idea of the estimates of these parameters. Therefore, Appendix B explains the estimation method and shows the estimated parameters.

7.2. Experimental Conditions

In this section, we will consider experimental conditions. In the data generating process, there are many parameters to specify. The parameters can be divided into three groups. First, the parameters of the panel data, that is, the number of individuals (N), the number of trials (K), and the number of different stimuli (R) can be varied. Then, the parameters of the model for brain activity of Section 5.2 can be different. Those parameters are β, Σ_β, and σ_ε for both ERN and Pe. At last, the parameters of Equation (4), that is, γ, δ, ρ, and σ_ϵ, can be changed.

Varying all parameters would lead to a very large simulation study, too large for the present study. Therefore, we will choose some parameters to have the same value during the simulations. To make the simulation study as realistic as possible, the parameters that will not vary in the simulation, are set as close as possible to (estimates of) parameters of the empirical data.

First, the number of trials and the number of different stimuli will be the same as in the empirical data, so $K = 400$ and $R = 4$. Further, the covariance matrices for β (for ERN and Pe) will be set roughly the same as the estimates

in Section 5.2. Also, the parameters γ, δ, and ρ, of the model in Equation (4) will not vary in the simulation. All those parameters will be based on estimates in Appendix B, such that $\gamma = 0.3$, $\delta = [-3, -1.5, -3, -1.5]'$, and $\rho = -1.2$. Finally, setting $\sigma_\epsilon = 1$ ensures that the probability process has sufficient errors in trials in the end.

The remaining parameters are N, β, and σ_ε. We will explain why those parameters are valuable to change and which values they will have.

7.2.1. Parameter N

It is interesting to consider multiple values for N, as papers in the present literature also have different sizes of data sets. Besides, if the simulation study shows that changing the number of participants will not differ the value of Cronbach's α, money can be saved in future research by using fewer participants. We consider a low number of observations ($N = 25$), a normal number of observations ($N = 50$), and a, in the field of psychiatry, large number of observations ($N = 100$).

7.2.2. Parameter β

The main research goal is to find out whether the OH-method gives bias in the value of Cronbach's α, and, thus, in the number of errors needed for internal consistency. The assumption of Cronbach's α is that it is computed over the same items. Now, if the ERN and Pe values across different stimuli differ from each other, the items over which α is computed are clearly not the same. However, if those brain activity values do not differ for different stimuli, they can be interpreted as constituting the same item. Hence, different values for β are considered. However, the constant, that is, the first element in β, will be the same in each case. We set this constant roughly equal to the estimates of the logarithm of ERN and Pe values in Section 5.2. Now, in the first case, we will assume no difference in the ERN or Pe values over the different stimuli, so $\beta = [2, 0, 0, 0]'$. In the second case, we will assume that all stimuli give different brain activity values, or, $\beta = [2, 2, 4, 6]'$. The last case will take into account different ERN and Pe values for congruent and incongruent stimuli, but within congruent or incongruent stimuli there will be no difference, such that $\beta = [2, 2, 0, 2]'$. Note that, Section 5.2 shows that estimate of β for the empirical data mostly matches $\beta = [2, 0, 0, 0]'$.

7.2.3. Parameter σ_ε

In Section 4, it is described that the Signal-to-Noise ratio for the empirical data is low, such that there is much noise. However, improving technology in the future can increase this Signal-to-Noise ratio. Therefore, we choose a high value for the variance ($\sigma_\varepsilon = 0.75$) and a low value ($\sigma_\varepsilon = 0.05$).

All combinations of different parameters will give $3 \times 3 \times 2 = 18$ parameter sets. Those different parameter sets are shown in Table 11.

Table 11: All parameter sets with the corresponding varying parameters.

Parameter set	N	β	σ_ε
1	25	$[2, 0, 0, 0]'$	0.05
2	25	$[2, 0, 0, 0]'$	0.75
3	25	$[2, 2, 4, 6]'$	0.05
4	25	$[2, 2, 4, 6]'$	0.75
5	25	$[2, 2, 0, 2]'$	0.05
6	25	$[2, 2, 0, 2]'$	0.75
7	50	$[2, 0, 0, 0]'$	0.05
8	50	$[2, 0, 0, 0]'$	0.75
9	50	$[2, 2, 4, 6]'$	0.05
10	50	$[2, 2, 4, 6]'$	0.75
11	50	$[2, 2, 0, 2]'$	0.05
12	50	$[2, 2, 0, 2]'$	0.75
13	100	$[2, 0, 0, 0]'$	0.05
14	100	$[2, 0, 0, 0]'$	0.75
15	100	$[2, 2, 4, 6]'$	0.05
16	100	$[2, 2, 4, 6]'$	0.75
17	100	$[2, 2, 0, 2]'$	0.05
18	100	$[2, 2, 0, 2]'$	0.75

7.3. Simulation Outcomes

The simulation study gives two matrices for both ERN and Pe, namely, Y and X. Below, we will discuss results that can be obtained using these matrices. Note that we have those results for each parameter set of Table 11.

7.3.1. Simulation: Outcome A

To get an idea of α_T, Y can be used. However, computing α_T over all trials will probably result in an α that is close to one (Eggen and Sanders, 1993, p. 44). Therefore, it will be more interesting to consider α_T for a part of the trials, namely, for $k \in \{2, 3, ..., 28\}$

7.3.2. Simulation: Outcome B

Next to α_T, we also consider α_{OH} using matrix X. The value of α_{OH} will be computed based on k errors, with $k \in \{2, 3, ..., 14\}$.

7.3.3. Simulation: Outcome C

The most important goal of this paper is to investigate possible bias in the OH-method. Therefore, we will also compare α_T with α_{OH}. The differences for those values $(\alpha_T - \alpha_{OH})$, computed in each run of the simulation, can be used to obtain the summary statistics of the histograms for this difference for different values of k.

7.3.4. Simulation: Outcome D

Next to the actual values of α and the results for the main goal of this paper, influences of the varying parameters on Cronbach's α can be considered. To find influential parameters, all values of α, coming from different parameter sets, are regressed on a constant and dummies for the varying parameters, such that we explore an analysis of variance, that is,

$$\alpha = \theta_0 + \theta_1 D_{N=25} + \theta_2 D_{N=50} + \theta_3 D_{\beta=[2,0,0,0]'} + \theta_4 D_{\beta=[2,2,4,6]'} + \theta_5 D_{\sigma_\varepsilon=0.05} + \eta, \tag{5}$$

where D indicates a dummy, and where η is assumed to be normally distributed. Here, α can refer to α_T or α_{OH} and α can be based on different values for k.

7.3.5. Simulation: Outcome E

Another way to decide upon internally consistent ERN and Pe values is using test-retest reliability. This estimate of reliability does not bring the same problems with it as Cronbach's α does, as there is no need to have the same items. There is a possibility that this way of computing reliability will give better results. Test-retest reliability is computed as follows if the 'complete' matrix Y is considered. The matrix is divided in two $N \times \frac{K}{2}$ matrices. For both matrices, the mean over the rows is computed. Then, the correlation of those mean vectors is computed. This correlation is called the test-retest reliability. If this reliability value is high, the measurements in the two matrices are comparable and thus reliable. Computing the test-retest reliability from matrix Y gives the 'true' test-retest reliability.

However, for computing test-retest reliability from the empirical data, we need to take into account that the matrix X contains missings. To compute test-retest reliability for X, we again divide the matrix in two $N \times \frac{K}{2}$ matrices. Then, we can also consider the mean of the rows, but now, we should not take into account the zeros. Correlating the mean vectors gives test-retest reliability.

7.3.6. Simulation: Outcome F

To find a variant on the OH-method that can be more reliable, we use lower bounds of the distributions in Method C, where we (partly) assumed $\mathcal{A}.3$, of Section 6. To do so, Method C is repeated for the simulated data and the lower bound of the confidence interval is computed in each run. If these lower bounds come close to α_T, they can be a good estimate of α in future research.

7.4. Simulation Set Up

Now, we will shortly explain the simulation set up and some details. There are 18 possible parameter sets. The steps in the simulation for one such a parameter set are shown below.

1. Set the parameters.
2. Create the exploratory variables.
3. Run the model for brain activity to obtain Y.
4. Run the model for finding the correct trials to obtain P^*.
5. Obtain X by element wise multiplying Y with P^*.
6. Use matrices X and Y to obtain the different outcomes as described above.

We will iterate S times through steps 3 till 6, where S is equal to 100. There are some details to note. First, next to random error processes (ε and ϵ), we also take a draw for K_i in each run based on the histogram in Figure 4a. Also, p_{i1} will be drawn of a uniform distribution (U(0,1)) in each run of the simulation. We will also change the randomisation of the stimuli (matrix D in the DGP) in each run.

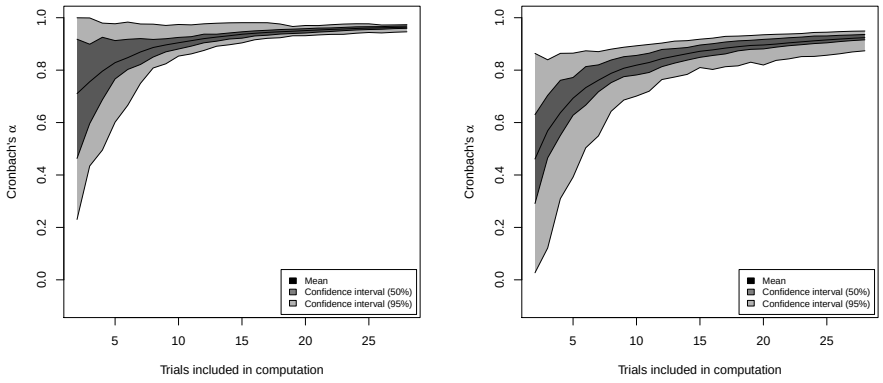
7.5. Results

In this section, we consider results of the simulation study. Only results for the ERN values are considered, as outcomes for ERN and Pe are similar. As parameter estimation in Section 5.2 shows that parameter set 1 and 2 are most closely to the empirical data, we will mainly show results for those parameter sets. Therefore results for some parameter sets are not shown. Confidence intervals shown in this section are empirical, that is, decided by sorting a certain vector and taking a predefined percentage of the middle values to be confident. We discuss the results for each of the outcomes above.

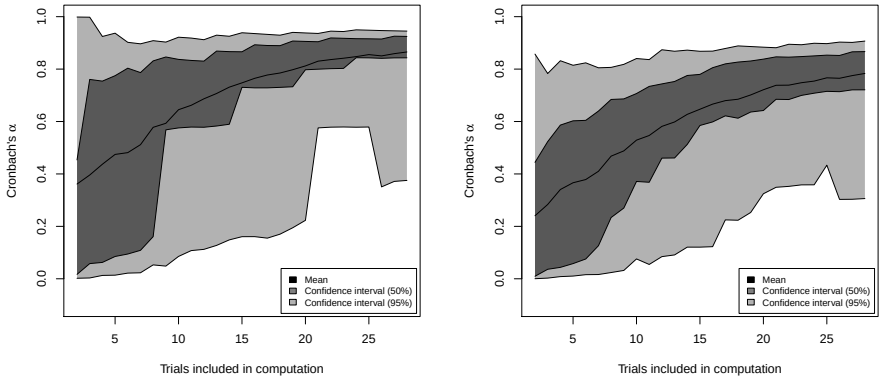
7.5.1. Simulation: Outcome A

In Figure 10, α_T is shown for parameter set 1, 2, 3, and 4 for $k \in \{2, 3, ..., 28\}$. The figures show that the value of β and the value of σ_ε have impact. Taking $\beta = [2, 2, 4, 6]'$ gives much more uncertainty around α_T and changing σ_ε from 0.05 to 0.75 shows that Cronbach's α increases more slowly.

The uncertainty shown in the figures is cannot be due to violation of assumptions $\mathcal{A}.1$, $\mathcal{A}.2$, and $\mathcal{A}.3$ of Cronbach's α because it is ensured that, for α_T, trials in the same columns corresponds to each other. Therefore, the uncertainty comes from having different sets of errors, that is, different runs in the simulation. This implies that taking one particular set of errors (for example, taking the first fourteen errors) influences the results. Note that there is more uncertainty in case of different stimuli causing different brain activity values ($\beta = [2, 2, 4, 6]'$). Nevertheless, this could also not be due to violation of the assumption.



(a) $N = 25$, $\beta = [2, 0, 0, 0]'$, $\sigma_\varepsilon = 0.05$ (b) $N = 25$, $\beta = [2, 0, 0, 0]'$, $\sigma_\varepsilon = 0.75$



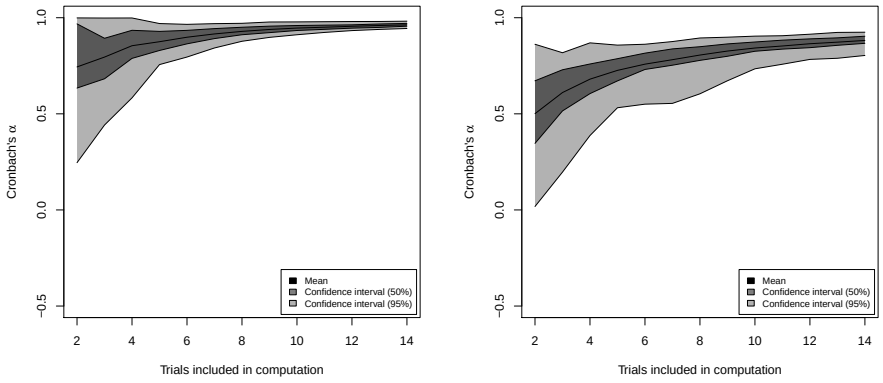
(c) $N = 25$, $\beta = [2, 2, 4, 6]'$, $\sigma_\varepsilon = 0.05$ (d) $N = 25$, $\beta = [2, 2, 4, 6]'$, $\sigma_\varepsilon = 0.75$

Figure 10: α_T, ERN

7.5.2. Simulation: Outcome B

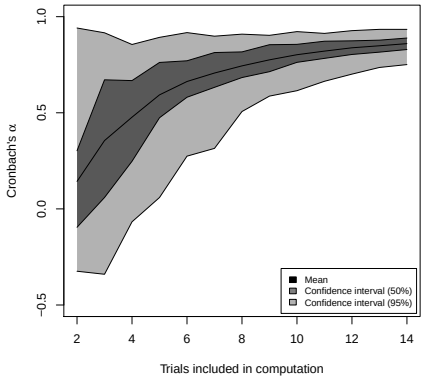
In Figure 11, α_{OH} is shown for parameter set 1, 2, 3, and 4 for $k \in \{2, 3, ..., 14\}$. These figures show comparable results as the figures in Figure 10.

As violation of assumption $\mathcal{A}.1$, $\mathcal{A}.2$, and $\mathcal{A}.3$ of Cronbach's α could not be the reason for the uncertainty in α_T, it is probably also not the reason for the uncertainty in α_{OH}. Nevertheless, it can have little influence as there is somewhat more uncertainty for lower k. But generally, the fact that figures of α_T and α_{OH} are comparable shows that the uncertainty is probably not due to the violation of the assumptions of Cronbach's α, but due to the specific set of errors that is considered.

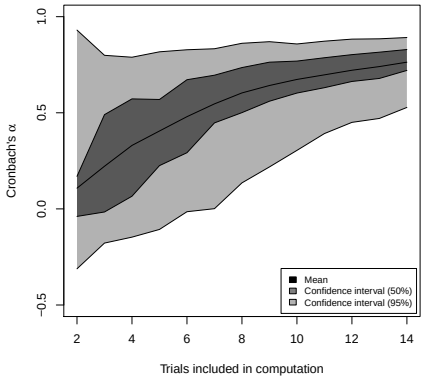


(a) $N = 25$, $\beta = [2, 0, 0, 0]'$, $\sigma_\varepsilon = 0.05$

(b) $N = 25$, $\beta = [2, 0, 0, 0]'$, $\sigma_\varepsilon = 0.75$



(c) $N = 25$, $\beta = [2, 2, 4, 6]'$, $\sigma_\varepsilon = 0.05$



(d) $N = 25$, $\beta = [2, 2, 4, 6]'$, $\sigma_\varepsilon = 0.75$

Figure 11: α_{OH}, ERN

7.5.3. Simulation: Outcome C

In Table 12, the median, lower bound, and upper bound of the histograms for the differences $\alpha_T - \alpha_{OH}$ are shown for $k = 2$, $k = 8$ and $k = 14$ error trials. Histograms are not shown, but they are bell shaped. The table is based on $N = 25$, but including more participants ensures in most histograms that differences are slightly smaller.

Now, it can be decided whether there is bias in the value of α. There are two ways to look at this issue. First, the table shows that the medians of all histograms are close to zero, meaning that in most cases, α_T and α_{OH} do not differ substantially. So, considering the median, we could say that there is no bias in the values for Cronbach's α, and thus there is no bias in the value of k^*. However, as a second way to look at this issue, confidence intervals can be

considered. We see that confidence intervals are generally wide, such that there are significant differences between α_T and α_{OH}. The size of this bias depends on the value of k and the parameter set that is considered. For example, confidence intervals become much wider if we change β from $[2,0,0,0]'$, to $[2,2,4,6]'$. More important to note is that, if k increases, the confidence intervals become smaller.

Table 12: The median, lower bound and upper bound of the simulated histograms for $\alpha_T - \alpha_{OH}$ for ERN.

N	β	σ_ε	k	Median	Lower bound	Upper bound
25	$[2,0,0,0]'$	0.05	2	0.00	-0.76	0.50
25	$[2,0,0,0]'$	0.75	2	-0.04	-0.52	0.43
25	$[2,2,4,6]'$	0.05	2	0.16	-0.88	1.14
25	$[2,2,4,6]'$	0.75	2	0.07	-0.69	0.90
25	$[2,0,0,0]'$	0.05	8	-0.04	-0.14	0.05
25	$[2,0,0,0]'$	0.75	8	-0.02	-0.19	0.17
25	$[2,2,4,6]'$	0.05	8	-0.08	-0.75	0.25
25	$[2,2,4,6]'$	0.75	8	-0.12	-0.70	0.33
25	$[2,0,0,0]'$	0.05	14	-0.03	-0.07	0.01
25	$[2,0,0,0]'$	0.75	14	-0.02	-0.10	0.07
25	$[2,2,4,6]'$	0.05	14	-0.08	-0.72	0.08
25	$[2,2,4,6]'$	0.75	14	-0.10	-0.70	0.17

Wide confidence intervals show that α_{OH} is an unreliable estimate for α_T. Large differences between α_T and α_{OH}, (e.g. $\alpha_T - \alpha_{OH} = -0.80$) mean that α_T can be very small (e.g. 0.1), indicating that the corresponding value of k is far too low, while α_{OH} can, for the same value of k, be close to one (e.g. 0.9), indicating that this value of k is perfect. In this case, using α_{OH} as an estimate of α_T can seriously bias in the number of errors needed for internal consistency.

To decide upon the number of error trials, one could simulate those histograms for the parameter set that corresponds to their empirical data. Then, they could decide the width of the confidence intervals for different values of k. If intervals are small enough, α_{OH} is a good estimate for α_T, so that the number of errors coming from the OH-method is reliable.

Remark that, if k increases, the upper bounds of the confidence intervals decrease fast towards zero, while lower bounds increase slowly to zero. This means that, for higher values of k, the difference $\alpha_T - \alpha_{OH}$ is very low on the positive side, but can be high on the negative side. Therefore, overestimation of α_T can be large, but underestimation is often very small, so overestimation happens more extreme than underestimation. If α_T is overestimated by α_{OH}, the number of errors taken to have internal consistency is underestimated. Therefore, an underestimation of k will happen more often, especially when k increases.

In sum, considering the median, α_{OH} is a reliable estimate of α_T, such that there is no bias in the number of errors concluded. However, looking at the confidence intervals, there can be significant bias. The size of the bias depends on the parameters set and the value of k. If k increases, α_{OH} becomes a more

reliable estimate of α_T, such that bias will be lower. Finally, α_{OH} more often overestimates α_T.

7.5.4. Simulation: Outcome D

To find the factors influencing Cronbach's α, the regression of Equation (5) is performed. Table 13 shows the estimates of the coefficients and the corresponding p-values. To decide which factors significantly contributes to values of α, a one percent significance level is used. First, the constants (θ_0) are increasing with the value of k and the value of α_{OH} is, on average, larger than that of α_T, indicating an overestimation. Furthermore, most p-values for θ_1 and θ_2 are not significant. This indicates that the number of observations N does not significantly influence the value of α. Therefore, having 25 participants will give, on average, the same value of α as having 50 or 100 participants. This means that, if one is only interested in Cronbach's α, money can be saved by taking less observations. In addition, θ_3 and θ_4 are significant for every case. With $\beta = [2, 2, 0, 2]'$ as baseline, α increases if we change β to $[2, 0, 0, 0]'$ and α decreases if we change β to $[2, 2, 4, 6]'$. Therefore, if we want to use a simulation study to decide upon the number of errors needed for internal consistency, it is important to know whether different stimuli causes brain activity values to differ. We can test this with the method of Section 5.2. At last, all coefficients θ_5 are significant with positive coefficients. This means that, on average, α increases if we decrease σ_ε.

Table 13: The estimates of the coefficients (Est) and the corresponding p-values (p) for Equation (5).

			N		β		σ_ε
		constant	25	50	$[2,0,0,0]'$	$[2,2,4,6]'$	0.05
		(θ_0)	(θ_1)	(θ_2)	(θ_3)	(θ_4)	(θ_5)
α_T,	Est	0.313	-0.003	0.007	0.201	-0.122	0.201
$k = 2$	p	0.000	0.862	0.661	0.000	0.000	0.000
α_T,	Est	0.664	0.000	0.024	0.111	-0.217	0.150
$k = 8$	p	0.000	0.959	0.055	0.000	0.000	0.000
α_T,	Est	0.769	0.002	0.016	0.075	-0.141	0.121
$k = 14$	p	0.000	0.710	0.014	0.000	0.000	0.000
α_{OH},	Est	0.389	-0.032	0.011	0.164	-0.299	0.213
$k = 2$	p	0.000	0.042	0.505	0.000	0.000	0.000
α_{OH},	Est	0.778	-0.012	0.003	0.035	-0.164	0.143
$k = 8$	p	0.000	0.014	0.618	0.000	0.000	0.000
α_{OH},	Est	0.863	-0.006	-0.001	0.021	-0.101	0.097
$k = 14$	p	0.000	0.031	0.604	0.000	0.000	0.000

7.5.5. Simulation: Outcome E

Outcomes of test-retest reliability are comparable with outcomes of Cronbach's α. Therefore, the results are not shown here. Also in test-retest reliability confidence intervals are wide. However, comparing test-retest reliability with

Cronbach's α over the parameter sets shows that test-retest reliability is most of the times somewhat higher. Nevertheless, it is important that results coming from the empirical data (matrix X) look like results from the simulated 'complete' data (Y) so that the 'empirical data' method gives a good estimate of the true reliability. Still, this is not the case as results for the empirical data show much larger confidence intervals than results for the 'true' method. Therefore, using test-retest reliability will not improve upon using Cronbach's α.

7.5.6. Simulation: Outcome F

Section 7.5.3 showed that α_{OH} mostly overestimates α_T. However, Section 6 showed that α_{OH} often underestimates α_M. Therefore, the relation $\alpha_T \leq \alpha_{OH} \leq \alpha_M$ approximately holds. Consequently, to get an estimate of α_T using the empirical data, we can consider lower bounds of the empirical distributions. Table 14 shows the percentages that α_T is higher than the lower bound of the simulated distributions using Method C (where we partially assumed $\mathcal{A}.3$) of Section 6. We see that those percentages are quite high if k is small. Therefore, taking the lower bound of the empirical distribution makes sure that we, with small values for k, at least underestimate α_T instead of overestimating it. Underestimating the truth ensures that more errors are taken to achieve internal consistency. So, underestimation makes the number of errors that are reported too safe instead of not safe enough as is the case with overestimating. As we mostly overestimated α_T using α_{OH} and as we underestimate it with those lower bounds, we recommend to use lower bounds of the empirical distributions of Cronbach's α to estimate α_T. This can give us too many error trials, but having too many is less worse than having not enough errors. Note that the empirical distributions of Method C is based on violation of assumptions $\mathcal{A}.1$ and $\mathcal{A}.2$ of Cronbach's α. However, results show that this violation has probably no influence on the value of α, such that the lower bounds are reliable estimates.

Table 14: The percentage of times that the lower bound of the simulated empirical distributions using Method C of Section 6 is lower than α_T for different values of k and for parameter sets 1 to 6.

Parameter	k												
set	2	3	4	5	6	7	8	9	10	11	12	13	14
1	85	77	68	65	57	45	45	43	34	37	34	25	21
2	87	91	88	85	83	82	81	79	74	70	71	68	70
3	95	85	79	76	62	63	62	51	52	51	47	47	47
4	100	97	95	88	81	78	73	66	62	60	58	59	61
5	70	54	44	41	33	25	24	13	15	9	9	5	5
6	97	81	80	75	69	65	61	54	49	49	51	43	40

7.6. Conclusions

The goal of the simulation study was to consider whether the problems as defined in Section 3.2 really cause bias in Cronbach's α computed using the OH-

method. The simulation is also used to find which factors influences Cronbach's α and to consider other methods for computing internal consistency. The most important conclusions are summarised below.

First, α_T gives sizable confidence intervals. Therefore, α_T itself already contains a lot of uncertainty. However, in the computation of α_T, assumptions $\mathcal{A}.1$, $\mathcal{A}.2$, and $\mathcal{A}.3$ are fulfilled. Therefore, the uncertainty needs to come from the fact that different sets of errors are taken in each run of the simulation.

Secondly, α_{OH} shows results comparable with the results of α_T. This would suggest that violating $\mathcal{A}.1$ and $\mathcal{A}.2$ does not really influence the results of the OH-method. However, it suggests again that, the particular set of errors has influence.

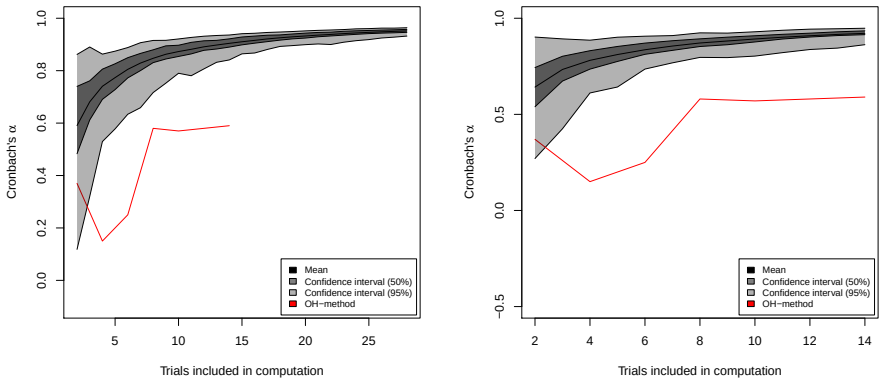
Furthermore, considering the median of $\alpha_T - \alpha_{OH}$, there is no bias in the values of α. However, confidence intervals are wide, showing that, especially when k is small, α_{OH} is not a reliable estimate for α_T. However, as k increases, the estimate α_{OH} becomes more reliable. Additionally, if α_{OH} is an unreliable estimate of α_T it is often the case that α_{OH} overestimates α_T such that the number of errors needed for internal consistency is underestimated. Therefore, in general, it can be the case that α_{OH} gives an underestimated number of errors (k^*). However, the higher this number of errors, the smaller the bias in α_{OH}.

Finally, as for small k confidence intervals are wide, we cannot use α_{OH} as a reliable estimate for α_T for smaller values of k. Nevertheless, for small values of k the lower bound for the empirical distributions of α is almost always underestimating α_T. Underestimating is safer than overestimating (α_{OH} overestimates α_T). Therefore, we can use this lower bound as an estimate for α_T for small values of k and we can use α_{OH} as estimate for α_T if k increases (and confidence intervals become smaller). Note that, to infer conclusions from the confidence intervals, they need to be computed using parameter estimates from the empirical data in the simulation.

8. Recommendation

Using the conclusions of Sections 6 and 7, we could consider which number of errors (k^*) would be decided for the empirical data of Section 4. The conclusions give the following procedure to find k^*. First, we need to estimate the parameters of the DGP based on the empirical data, so that we can create an appropriate parameter set. For this parameter set, the parameter estimates of Section 5.2 and Appendix B are used. Using this parameter set, we can simulate lower and upper bounds of the difference $\alpha_T - \alpha_{OH}$. From these confidence intervals, we can decide the value of k for which α_{OH} becomes a reliable estimate of α_T. Then, the lower bounds (for smaller values of k) of the empirical distribution of Cronbach's α of Method C (see Table 10) can be used to at least underestimate α_T. If these lower bounds already show high values of α, we can find the corresponding number of errors to obtain k^*. If not, we consider α_{OH} for values above the k coming from the confidence intervals. As those values for α_{OH} are reliable, we can simply find an α that is high enough. We will only give a recommendation for ERN values as we mostly discussed those values.

First, the plots for α_T and α_{OH} for the simulated version of the empirical data are shown in Figure 12. Figure 12b shows that the simulated versions of α_{OH} are higher than the actual α_{OH}. Therefore, simulation of the empirical data gives an overestimated Cronbach's α in the OH-method. Accordingly, there is a high probability that α_T is also overestimated. Therefore, we cannot use α_T of Figure 12a to find the number of errors. However, as both α_T and α_{OH} are being overestimated, the differences $\alpha_T - \alpha_{OH}$ do probably give accurate results.



(a) α_T (b) α_{OH}

Figure 12: The values for Cronbach's α when using the simulated version of the empirical data together with Cronbach's α coming from the OH-method from the empirical data.

To find k^*, the first step is to find the value of k for which the confidence intervals of $\alpha_T - \alpha_{OH}$ are small, such that α_{OH} is a reliable estimate of α_T. Table 15 shows the lower and upper bounds for the histograms of $\alpha_T - \alpha_{OH}$ for different values of k. Suppose that we assume that estimates of α_T are reliable if uncertainty is smaller than 0.10. Then, the number of errors needed for α_{OH} to be a reliable estimate of α_T, based on a 95 percent confidence interval, is equal to ten (see Table 15). So, from $k = 10$ onwards, α_{OH} can be used to find the number of errors necessary to have internal consistency.

For all values of k smaller than ten, the lower bounds of the empirical distribution of α in Method C do probably give an underestimation of α_T, which is more safe than an overestimation. Table 10 shows that for values of k smaller than ten, the lower bound is smaller than 0.50. Therefore, we consider the values for α_{OH} in Table 8 for $k \geq 10$. We see that $\alpha_{OH} \geq 0.50$ when $k = 10$, indicating that ten errors will be enough to have internal consistency, or, that $k^* = 10$. However, note that, α_T can still be equal to 0.48 as the lower bound in Table 15 with $k = 10$ is -0.09. Furthermore, note that taking α equal to 0.50 is low. Many studies consider α to be high enough when it is at least 0.70. Doing

the procedure for this value of α would suggest that we need more than fourteen errors.

Table 15: The lower and upper bounds of the simulated histograms for $\alpha_T - \alpha_{OH}$ for ERN for the simulated version of the empirical data.

k	2	3	4	5	6	7	8
Lower bound	-0.54	-0.29	-0.21	-0.23	-0.16	-0.18	-0.14
Upper bound	0.30	0.14	0.12	0.14	0.12	0.10	0.06
k	9	10	11	12	13	14	
Lower bound	-0.14	-0.09	-0.10	-0.09	-0.08	-0.08	
Upper bound	0.07	0.08	0.08	0.07	0.06	0.05	

The number of errors needed to have internal consistent ERPs is equal to ten if we assume a maximum uncertainty of 0.10. This is more than eight errors which was concluded in the OH-method. Therefore, considering the particular empirical data set of Section 4, this recommended method suggest more errors needed for internal consistency than the OH-method does.

9. Conclusion and Discussion

Present studies relate average ERPs to psychological disorders. In those studies, average ERPS are often derived as the average ERPs over all error trials. However, if a participant only made a few errors, this average can be unreliable. Therefore, Olvet and Hajcak (2009), Marco-Pallares et al. (2011), Pontifex et al. (2010), Meyer et al. (2013), Rietdijk et al. (2014) used the OH-method to find the number of errors that makes ERPs internal consistent. This number of errors (k^*) is used as a selection criterion and average ERPs are only derived for participants that failed more than k^* times.

The main purpose of this research was to examine the OH-method. Namely, the OH-method ignores many error trials and it violates the assumptions of Cronbach's α. Therefore, it is possible that the number of errors (k^*) coming from the OH-method is biased. Secondly, in the derivation of average ERPs, many studies use all error trials to compute an average. However, including all error trials only gives reliable ERPs if the brain activity values within a person are independent over the error trials. Therefore, another goal of this research was to justify averaging over all error trials.

We started with examining independency across brain activity within participants, as finding a selection criterion is useless if averaging all error trials is not justified. For this purpose, we divided independency in two issues. We considered whether ERPs change over time by testing for a trend and we decided whether different stimuli give different ERPs. For this last purpose, we modelled the ERPs using a random parameter model. We concluded no differences in brain activity values over time and for different stimuli. Therefore, brain activity within a participant is independent over the error trials, so that

averaging all error trials is justified. Thus, research on the selection criterion, or, on the OH-method, is useful.

To examine the OH-method, we first considered empirical distributions of Cronbach's α. It turned out that taking different sets of error trials can give completely different values of α. However, permuting the same set of error trials did not change the value of α in case that $k = K_m$, indicating that the specific set of errors matters but violation of the assumptions does not.

We also considered a simulation study, where we generated data for 18 different parameter sets. From this simulation, the following can be concluded. The violation the assumptions of Cronbach's α does not give bias in the values of α. However, the fact that specific sets of errors are considered does give bias. Furthermore, confidence intervals for small values of k, show that α_{OH} is an uncertain estimator of α_T. The size of this uncertainty can cause serious bias in the number of errors that is needed to have internal consistent ERPs. However, when increasing k, the confidence intervals become smaller, indicating less bias, or, in other words, showing that α_{OH} becomes a more reliable estimate for α_T.

In sum, deriving ERPs by averaging all error trials is justified. Therefore, it is important to find the number of errors that makes this ERPs internal consistent. For finding this number of errors, the OH-method can be used if the two problems arising from this method do not bias the results. First, violation of the assumptions of Cronbach's α does not give much bias. However, the particular selection of errors used to compute Cronbach's α can bias the value of α. Therefore, the number of errors coming from the OH-method can be biased because of a selection of errors is used to compute Cronbach's α. Nevertheless, this bias only occurs for smaller values of k.

The simulation study also gave another method to find the number of errors that is needed for internal consistency. Applying this method to the empirical data shows that 10 errors are needed for internal consistency, instead of 8 as decided in the OH-method. However, this result is based on the fact that having $\alpha \geq 0.50$ would be high enough.

This research contains some doubts that can be thought of. First, ERPs are measured when it is observed that participants make errors. Now, it can well be the case that a participant is not aware of making an error. Then, the brain of this participant do not show ERPs. However, we will use these errors as if they showed real ERPs. This can create bias in the average ERP. Therefore, it can be interesting for future research to also ask participants about the consciousness of making an error. Secondly, we saw in Section 8 that α_{OH} based on simulated data (using parameter estimates of the model of Section 5.2) overestimates α_{OH} based on empirical data. This would suggest that the model of Section 5.2 does not exactly generate the empirical data. This can be due to the fact that the assumptions underlying this model are not correct. More research on correct models is preferable. Also, in practice, most researchers advise at least a value of α that is higher than 0.70. In this case, the empirical data showed that $k^* = 26$ for the OH-method. However, to obtain this k^*, many participants were excluded because of not having at least 26 errors. Therefore, to decide whether α_{OH} is reliable in this case, simulated data based on a parameter set

with correct parameters can be performed. Further, as the main purpose of most studies is to derive reliable ERPs, future research can be developing an EEG system where average ERPs are directly measured. Then, when averages do not differ anymore, the participant can exit the experiment. In this way, all participants give usable average ERPs and none of the errors is ignored. Besides, problems coming from using Cronbach's α are avoided.

References

Pearl Chiu and Patricia Deldin. Neural evidence for enhanced error detection in major depressive disorder. *American Journal of Psychiatry*, 164(4):608–616, 2007.

Lee J Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334, 1951.

T.J.H.M. Eggen and P.F. Sanders. *Psychometrie in de praktijk*. Cito Instituut voor Toetsontwikkeling, 1993. ISBN 9789090065083. URL `http://books.google.nl/books?id=xmlKAAAACAAJ`.

Barbara A Eriksen and Charles W Eriksen. Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & psychophysics*, 16(1):143–149, 1974.

William J Gehring, Joseph Himle, and Laura G Nisenson. Action-monitoring dysfunction in obsessive-compulsive disorder. *Psychological science*, 11(1): 1–6, 2000.

Greg Hajcak. What we've learned from mistakes insights from error-related brain activity. *Current Directions in Psychological Science*, 21(2):101–106, 2012.

Avram J Holmes and Diego A Pizzagalli. Spatiotemporal dynamics of error processing dysfunctions in major depressive disorder. *Archives of General Psychiatry*, 65(2):179–188, 2008.

Sönke Johannes, Bernardina M Wieringa, Wido Nager, Dominik Rada, Reinhard Dengler, Hinderk M Emrich, Thomas F Münte, and Detlef E Dietrich. Discrepant target detection and action monitoring in obsessive–compulsive disorder. *Psychiatry Research: Neuroimaging*, 108(2):101–110, 2001.

Richard Arnold Johnson, Dean W Wichern, and Pearson Education. *Applied multivariate statistical analysis*, volume 4. Prentice hall Englewood Cliffs, NJ, 1992.

Cecile D Ladouceur, Ronald E Dahl, Boris Birmaher, David A Axelson, and Neal D Ryan. Increased error-related negativity (ern) in childhood anxiety disorders: Erp and source localization. *Journal of Child Psychology and Psychiatry*, 47(10):1073–1082, 2006.

Josep Marco-Pallares, David Cucurell, Thomas F Münte, Nadine Strien, and Antoni Rodriguez-Fornells. On the number of trials needed for a stable feedback-related negativity. *Psychophysiology*, 48(6):852–860, 2011.

Alexandria Meyer, Anja Riesel, and Greg Hajcak Proudfit. Reliability of the ern across multiple tasks as a function of increasing errors. *Psychophysiology*, 50(12):1220–1225, 2013.

Doreen M Olvet and Greg Hajcak. The stability of error-related brain activity with increasing trials. *Psychophysiology*, 46(5):957–961, 2009.

Matthew B Pontifex, Mark R Scudder, Michael L Brown, Kevin C O'Leary, Chien-Ting Wu, Jason R Themanson, and Charles H Hillman. On the number of trials necessary for stabilization of error-related brain activity across the life span. *Psychophysiology*, 47(4):767–773, 2010.

Wim JR Rietdijk, Ingmar HA Franken, and A Roy Thurik. Internal consistency of event-related potentials associated with cognitive control: N2/p3 and ern/pe. *PloS one*, 9(7):e102672, 2014.

Paravastu AVB Swamy. Efficient inference in a random coefficient regression model. *Econometrica: Journal of the Econometric Society*, pages 311–323, 1970.

Digby Tantam. Psychological disorder in adolescents and adults with asperger syndrome. *Autism*, 4(1):47–62, 2000.

Blair Wheaton. The sociogenesis of psychological disorder: An attributional theory. *Journal of Health and Social Behavior*, pages 100–124, 1980.

Nicola M Wöstmann, Désirée S Aichert, Anna Costa, Katya Rubia, Hans-Jürgen Möller, and Ulrich Ettinger. Reliability and plasticity of response inhibition and interference control. *Brain and cognition*, 81(1):82–94, 2013.

Appendix A. Derivation of Cronbach's α Being the Lower Bound of Reliability

In this appendix, we will derive the fact that Cronbach's α is the lower bound of the reliability following Eggen and Sanders (1993). Note that notation differs from notation used throughout the paper. Cronbach's α measures internal consistency of items. Items can be, for example, questions in surveys. Usually, there are multiple respondents for those items. That is, we measure an item score x_{ij} for respondent i in item j. However, mostly there is noise in the measurements. Therefore, the observed score can be decomposed as

$$x_{ij} = t_{ij} + e_{ij},$$

where t_{ij} is the true item score and e_{ij} is the measurement error which is uncorrelated with itself and the true scores and which has expectation zero. In repeated measures, a constant value of t_{ij} is expected within a respondent. To measure the reliability of the observed values, the following measure is used

$$\rho_{xt}^2 = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_e^2}. \tag{A.1}$$

Now, the observed values are reliable when ρ_{xt}^2 is close to one, or, in other words, when there is no measurement error ($\sigma_e^2 \approx 0$). The problem is that Equation (A.1) contains unknowns. Therefore, the reliability needs to be estimated. For this purpose, we use the following fact

$$\rho_{xt}^2 \geq \frac{K}{K-1}\left(1 - \frac{\sum_{j=1}^{K} \sigma_{x_{.j}}^2}{\sigma_x^2}\right),$$

where $\sigma_{x_{.j}}^2$ is the variance of the j^{th} question over all respondents and σ_x^2 is the variance of the total scores for each participant. The sample version of the right hand side is exactly Cronbach's α and this is a lower bound for the reliability of the observed scores. To obtain this inequality, we start with considering the variance of the difference between two true scores (for two items a and b)

$$\sigma_{t_{.a}-t_{.b}}^2 = \sigma_{t_{.a}}^2 + \sigma_{t_{.b}}^2 - 2\text{cov}(t_{.a}, t_{.b}) \geq 0,$$

where, for example, $t_{.a}$ can be seen as true scores for item a over all respondents. Now, we rewrite this as

$$\sum_{a \neq b} \left(\sigma_{t_{.a}}^2 + \sigma_{t_{.b}}^2\right) \geq 2 \sum_{a \neq b} \text{cov}(t_{.a}, t_{.b}). \tag{A.2}$$

We also consider the sum, where the same pairs of items are allowed, or,

$$\begin{aligned}
\sum_a \sum_b \left(\sigma_{t_{.a}}^2 + \sigma_{t_{.b}}^2\right) &= 2K \sum_a \sigma_{t_{.a}}^2 \\
&= 2 \sum_a \sigma_{t_{.a}}^2 + \sum_{a \neq b} \left(\sigma_{t_{.a}}^2 + \sigma_{t_{.b}}^2\right) \\
&\geq 2 \sum_a \sigma_{t_{.a}}^2 + 2 \sum_{a \neq b} \text{cov}(t_{.a}, t_{.b}),
\end{aligned}$$

where the inequality comes from Equation (A.2). Looking at those equations, we can write

$$(K-1)\sum_a \sigma^2_{t_{.a}} \geq \sum_{a\neq b} \mathrm{cov}(t_{.a}, t_{.b}).$$

Furthermore, we can write the 'true' variance as

$$\begin{aligned}\sigma^2_t &= \sigma^2(\sum_a t_{.a}) \\ &= \sum_a \sigma^2_{t_{.a}} + \sum_{a\neq b} \mathrm{cov}(t_{.a}, t_{.b}) \\ &\geq \frac{K}{K-1} \sum_{a\neq b} \mathrm{cov}(t_{.a}, t_{.b}),\end{aligned}$$

where, using the fact that the measurement error is uncorrelated with itself and the true score, the last term can be written as

$$\sum_{a\neq b} \mathrm{cov}(t_{.a}, t_{.b}) = \sum_{a\neq b} \mathrm{cov}(x_{.a}, x_{.b}) = \sigma^2_x - \sum_a \sigma^2_{x_{.a}}.$$

Now, we have

$$\begin{aligned}\rho^2_{xt} &= \frac{\sigma^2_t}{\sigma^2_x} \\ &\geq \frac{K}{K-1} \frac{\sigma^2_x - \sum_a \sigma^2_{x_{.a}}}{\sigma^2_x} \\ &= \frac{K}{K-1}\left(1 - \frac{\sum_a \sigma^2_{x_{.a}}}{\sigma^2_x}\right),\end{aligned}$$

such that we have a lower bound for the reliability. The sample version of this lower bound is called Cronbach's α, which contains quantities that can be estimated.

Furthermore, using this derivation, the range of Cronbach's α can be decided. Combining Equation (A.1) with the fact that variances are always non negative results in the fact that ρ^2_{xt} cannot exceed one. Now, as $\rho^2_{xt} \geq \alpha$, it is impossible that α will exceed one. Therefore, the upper bound of Cronbach's α is equal to one. However, Cronbach's α cannot be bounded from below such that $\alpha \in (-\infty, 1]$.

Appendix B. Estimation of the Parameters in the Model for Correct Trials

This section gives an impression of the parameters in Equation (4). First, we will focus on how to estimate the parameters in Equation (4). This is not straight forward as the variable $\sum_{l=j-4}^{j-1} p_{il}$ is in the model. Therefore, we need to estimate sequential, so that we can update this term each trial. For estimating the parameters, a Maximum Likelihood procedure is used. The probability of failure depends on a number of variables as described in Equation (4). To ensure probabilities as the outcome of the model, a logistic transformation is suitable. The log likelihood function is given by

$$\begin{aligned}
&\log\left(\prod_{i=1}^{N}\prod_{j=1}^{J} g\left(\gamma,\delta,\rho|p_{ij}^{*},d_j,\sum_{l=j-4}^{j-1} p_{il}\right)\right)\\
=&\sum_{i=1}^{N}\sum_{j=1}^{J} p_{ij}^{*}\log\left(F\left(\gamma z_j + d_j'\delta + \rho\sum_{l=j-4}^{j-1} p_{il}\right)\right)\\
+&\sum_{i=1}^{N}\sum_{j=1}^{J}\left(1-p_{ij}^{*}\right)\log\left(1-F\left(\gamma z_j + d_j'\delta + \rho\sum_{l=j-4}^{j-1} p_{il}\right)\right),
\end{aligned}$$

with J the number of trials that are used for estimation. Parameter estimates are obtained by maximizing this log likelihood function. The estimates for γ, δ and ρ are

$$\hat{\gamma} = \ 0.32 \ , \hat{\delta} = \begin{bmatrix} -2.89 \\ -1.65 \\ -2.84 \\ -1.52 \end{bmatrix}, \hat{\rho} = \ -1.21$$

First, $\hat{\gamma}$ is positive, indicating a higher probability of failing in the beginning of the experiment. Further, all values in $\hat{\delta}$ are negative. They correspond to the stimuli in the same order as in Table 1. From the values of $\hat{\delta}$ it can be seen that incongruent stimuli have a higher chance, that is, higher estimate, to become an error than the congruent stimuli. At last, $\hat{\rho}$ is negative, which implies that more errors in the current past, will give a smaller chance of making an error in the next iteration. This is not what we expected. However, in the empirical data, having an error followed by a correct response also occurs many times, such that ρ can be estimated negative.

Furthermore, the error process for the first participant is shown in Figure B.13. The line indicates whether a probability is high enough to be translated to a real error, so that, everything above the line is translated to an error. We see that many errors occur in the beginning. Those results change if an error term is added to the probability in the simulation. Then, errors also occur in the end of the experiment.

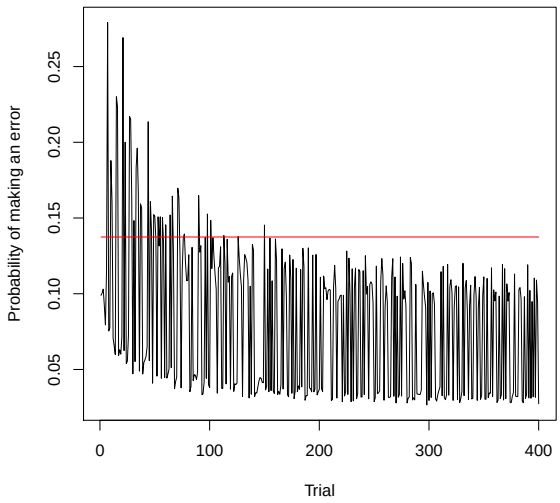


Figure B.13: The error process for the first participant.