International Journal of Psychophysiology 110 (2016) 56–65

Contents lists available at ScienceDirect

International Journal of Psychophysiology

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

The first independent study on the complex trial protocol version of the P300-based concealed information test: Corroboration of previous findings and highlights on vulnerabilities

Gáspár Lukács [a,1], Béla Weiss [b,1], Vera Daniella Dalos [a], Tünde Kilencz [a], Szabina Tudja [a], Gábor Csifcsák [a,c,*]

[a] Department of Cognitive and Neuropsychology, Institute of Psychology, Faculty of Arts, University of Szeged, Egyetem u. 2, 6722 Szeged, Hungary
[b] Brain Imaging Centre, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Magyar Tudósok körútja 2, 1117 Budapest, Hungary
[c] Department of Psychology, University of Tromsø, Huginbakken 32, 9037 Tromsø, Norway

ARTICLE INFO

Article history:
Received 2 May 2016
Received in revised form 12 October 2016
Accepted 13 October 2016
Available online 15 October 2016

Keywords:
Memory detection
Concealed information test
P300
Complex trial protocol
Countermeasure
Verification

ABSTRACT

More than a dozen studies of the Complex Trial Protocol (CTP) version of the P300-based Concealed Information Test have been published since its introduction (Rosenfeld et al., 2008), and it has been fairly consistently proven to provide high accuracy and strong resistance to countermeasures (Rosenfeld et al., 2013). However, no independent authors have verified these findings until now. In the present, first independent study, we corroborate the accuracy and countermeasure-resistance of the CTP, when the probe item (critical presented information, e.g., crime detail; P) vs. all irrelevant items (Iall) comparison is used for classifying participants as guilty or innocent, but we also show that the CTP is severely vulnerable to countermeasures, when the P vs. the irrelevant item with the largest P300 responses (Imax) comparison is used. This latter measure can be defeated by creating "oddball" items among the irrelevant items (through targeting them with covert responses), and thereby making their P300 responses statistically indistinguishable from those of the probe item. Practical implications are discussed.

1. Introduction

Undetected deception may have high costs in certain scenarios, for example in connection with legal cases or counterterrorism – however, meta-analyses show that humans, without special aid, are rarely able to reliably discriminate lies from the truth, most usually demonstrating judgment accuracies similar to pure chance (Bond and DePaulo, 2006; Hartwig and Bond, 2011; Kraut, 1980). Moreover, no significant individual differences can be found; neither experience nor training improves the accuracy of judgments (Bond and DePaulo, 2008; Meissner and Kassin, 2002). As a technological aid, the polygraph was invented just about a century ago, and is today widely used in many countries all over the world. While it provides higher than chance accuracy, it suffers from various limitations, including severe vulnerability to countermeasures (National Research Council, 2003).

1.1. The P300 as a tool for detecting concealed information

A prominent alternative under development is the P300-based deception detection that is based on analyzing neural activity recorded by electroencephalography (EEG). In an EEG examination, electrodes are placed on the scalp, through which electrical activity in the brain can be detected. The P300 is an event-related potential with a positive peak arising most prominently above the parietal lobe, beginning usually around 300 ms after stimulus presentation (review: Polich, 2007). It is typically obtained through the "oddball" paradigm: when presenting a random sequence of stimuli, an infrequent stimulus will evoke the P300 wave if it is task-relevant and requires an overt or covert response that is different from the rest of the stimuli (so-called "standards"). Importantly, the probability with which a stimulus occurs robustly influences the magnitude of the P300: infrequent salient stimuli evoke larger P300 waveforms – an effect which is considered to reflect the involvement of limited-capacity cognitive processes in the generation of the P300 (Polich, 2007). According to the influential context-updating theory of the P300 (Donchin, 1981; Donchin and Coles, 1988), this waveform represents the updating of stimulus representations in working memory, a process that is highly context-dependent, i.e., is influenced by both immediate stimulus history and task demands (previous knowledge, expectation, selective attention, etc.). Alternatively, the P300 has been linked to the formation of decisions, reflecting the gradual accumulation of evidence until a decision boundary is reached (O'Connell et al., 2012; Twomey et al., 2015). Finally, more recent accounts of the P300 emphasize the reactivation of previously established stimulus-response associations, a process that also depends on stimulus frequency (Verleger et al., 2014, 2015).

[*] Corresponding author at: Department of Cognitive and Neuropsychology, Institute of Psychology, Faculty of Arts, University of Szeged, Egyetem u. 2, 6722 Szeged, Hungary.
E-mail address: gaborcsifcsak@yahoo.co.uk (G. Csifcsák).
[1] The first two authors contributed equally to this work.

http://dx.doi.org/10.1016/j.ijpsycho.2016.10.010
0167-8760/© 2016 Elsevier B.V. All rights reserved.

The sensitivity of the P300 to stimulus context and task demands can be used in the Concealed Information Test (CIT), also known as the Guilty Knowledge Test, a deception detection method that is based on the recognition of a certain stimulus, for example a crime-relevant information, among other, irrelevant stimuli (Lykken, 1959; Verschuere et al., 2011; Verschuere and Meijer, 2014). In an often used example to describe the CIT, various items are sequentially presented to a murder suspect, any of which could be the murder weapon, for example: "gun," "knife," "rope," etc. One of these items is a *probe* item: the true murder weapon with which the actual crime was committed. All other items are conventionally called *irrelevant* items. In EEG studies pertinent to our study, the number of different irrelevant items typically ranges from four to eight, and each item (including the probe) is equally repeated for example 40 or 50 times, presented in a random sequence. It is assumed that the suspect will recognize the true murder weapon only if he/she has participated in the murder. The recognition of the true murder weapon, as a consequently salient item among other items, will result in a detectably larger average P300 response.

Numerous articles on this subject have been published since the first successful experiments starting from the late 1980s (mainly: Farwell and Donchin, 1991; Rosenfeld et al., 1988). The great majority of these studies have been conducted in the laboratory of J. P. Rosenfeld, where the Complex Trial Protocol (CTP) version of the P300-based CIT was also introduced, and has been used in more than a dozen studies by now (Rosenfeld et al., 2013, 2008). In most of these studies, the CTP has been consistently found to achieve the goals of its conception: to improve general accuracy, and more importantly, to resist countermeasures that were found to greatly reduce accuracy in previously used methods (Mertens and Allen, 2008; Rosenfeld et al., 2004). In a related review published in 2012, the necessity of an independent replication of these otherwise successful series of studies was already remarked (Ben-Shakhar, 2012), but, to the best of our knowledge, no such attempts were reported to date. Besides replicating some of the findings, the main purpose of our study was to provide an outside view through a reconsideration of the previous studies with an emphasis on the findings related to the resistance to countermeasures, which is considered to be a key feature of the CTP method, distinguishing it from other deception detection methods (Rosenfeld, 2011; Rosenfeld et al., 2013).

1.2. Uninvestigated effects of countermeasures

The most effective countermeasures against the P300-based CIT were found to be concealed responses (e.g., small physical movements or recalling a name of a person) that are assigned to specific irrelevant items, and executed when those items appear (Mertens and Allen, 2008; Rosenfeld et al., 2004). It has been reasoned that these covert responses make the corresponding irrelevant items relevant during the task, and thus, the probe item would not be the only relevant item anymore (Rosenfeld et al., 2004). Consequently, the "oddball" nature of the paradigm is weakened, resulting in reduced differences between the probe- and the irrelevant item-induced P300 responses – thereby increasing the chances for a guilty participant to be classified as innocent.

As several studies seemed to prove the CTP highly countermeasure-resistant (Hu et al., 2012; Labkovsky and Rosenfeld, 2012; Rosenfeld et al., 2008; Rosenfeld and Labkovsky, 2010; Winograd and Rosenfeld, 2011), more recent research focused on other areas (optimization of parameters, etc., Hu et al., 2013; Meixner and Rosenfeld, 2014; Rosenfeld et al., 2015a, 2015b; Winograd and Rosenfeld, 2014). However, these studies have been using two kinds of P300-based measurements for the classification of participants as guilty or innocent, namely, the "P vs. Iall," and the "P vs. Imax" measures – and the countermeasure effects on the P vs. Imax were not as thoroughly tested, as on the P vs. Iall measure. The original P vs. Iall measure has been regularly used for classification since the first P300-based CIT studies (Farwell and Donchin, 1991; Wasserman and Bockenholt, 1989), including all CTP articles (see Rosenfeld et al., 2013). The P vs. Imax measure was introduced along with the CTP as an alternative analysis method (Rosenfeld et al., 2008), and has been used and reported in subsequent studies (Meixner et al., 2009; Meixner and Rosenfeld, 2014, 2011; Rosenfeld and Labkovsky, 2010), but not in all of them (see Rosenfeld et al., 2013).

While the P vs. Iall measure compares the P300 responses to the probe (P) with the P300 responses to all irrelevant items (Iall), the P vs. Imax measure compares the P300 responses to the probe with the P300 responses to the one irrelevant item that has evoked the largest average P300 among all the irrelevant items (which is the "Imax"). The advantage of this measure – as it was argued by the authors (Rosenfeld et al., 2008) – is that it may be able to provide a higher specificity (i.e., less false positive classifications). In studies using both measures, they proved to provide very similar accuracies, although with the P vs. Imax measure indeed having, in general, a slightly higher specificity (Meixner et al., 2009; Meixner and Rosenfeld, 2014, 2011; Rosenfeld et al., 2008; Rosenfeld and Labkovsky, 2010), which could indicate that it is the preferable alternative.

Finding an item that evokes the largest P300 response can also have a very important practical use in itself. When the relevant detail (i.e., the probe item), is not exactly known, then a group of items can be shown to the suspect, out of which the one that evokes the largest P300 responses would be selected as a "presumed probe". For example, a terrorist attack is about to happen, but it is not exactly known in which city, or on which date the attack will take place, although there are several assumed possibilities. In this case, a suspected conspirator could be presented these assumed possibilities to determine which of them evokes the largest P300, and whether the P300 of this presumed probe is significantly larger than those of the other items. This is a scenario that Meixner and Rosenfeld (2011) tested in a mock-terrorism experiment with very good results (Meixner and Rosenfeld, 2011; and more details on the theory in Rosenfeld, 2011, p. 83).

Since the P vs. Imax measure takes into account the largest irrelevant P300 alone, it is considerably more vulnerable to an "outlier" irrelevant item that evokes larger P300 responses than the rest of the irrelevant items. In this regard, the Probe vs. Imax approach is not only more rigorous in classifying examinees as guilty (Rosenfeld et al., 2008), but, to some extent, it could also be sensitive to the use of a countermeasure technique, since an outlier can also be created voluntarily: beside the first "oddball," i.e. the probe item, one may create a secondary "oddball," through targeting an irrelevant item with a unique covert response, while still keeping the majority of irrelevant items comparatively regular. This scenario would create a special oddball paradigm, with two salient items, i.e. the probe and the targeted irrelevant item, against the majority of the other items (Katayama and Polich, 1999).

Two articles on the CTP have been published that have examined the effect of countermeasures (covert responses) against less than the half of the irrelevant items, but neither of these reported P vs. Imax measures (Hu et al., 2012; Labkovsky and Rosenfeld, 2012). However, one of these articles (Hu et al., 2012) did report that, in the case of 2 countered irrelevant items out of 8, the P300 responses to probe and countered irrelevant items were significantly larger than those of non-countered irrelevant items, while, in the cases of 4 and 6 countered irrelevant items out of 8, only the probe but not the countered irrelevant items evoked P300 responses significantly larger than non-countered irrelevant items (Hu et al., 2012, p. 88). Despite this observation, no further investigation was recounted in this direction.

Our hypothesis pertinent to our study was that when only a small group of irrelevant items are countered, at least one of them will tend to evoke a P300 that approximates the P300 to the probe, and consequently, the accuracy of the P vs. Imax measure will be significantly reduced. To provide a clear proof of this vulnerability, we used countermeasures with a few small modifications in order to enhance them.

1.3. Restructuring and simplifying countermeasures

Initial P300-based CIT methods included a designated target item among the irrelevant items, to which a different behavioral response (key press) had to be executed when it appeared. However, Rosenfeld et al. (2008) have reasoned that this task drains processing resources, diverting attention from the recognition of the probe item, and thus also reducing the P300 response to it. Therefore the CTP was devised so that the probe and irrelevant items all required the same response, a key press indicating merely that the participant saw the displayed item (Rosenfeld et al., 2008, pp. 906, 907). Additionally, to hold attention throughout the task, after each trial of displaying a probe or an irrelevant item, a simple secondary decision task was presented, with a rare target item requiring a button press different from the response to the non-target items. In most of the following studies, this decision task involved strings of five identical numbers, where the string of 11111 was the target, and strings of four other numbers (22222, 33333, 44444, and 55555) were non-targets. Each of all these stimuli is presented for 300 ms, appearing within 2 s after each other, in the typical trial structure. Thus, the CTP method may reduce cognitive load during the probe-irrelevant discrimination task, but, overall, the combined task is fairly demanding, and especially so if a participant tries to consistently execute a number of predefined countermeasures to various items. Our assumption here is that the CTP's resistance to countermeasures is at least partly due to this increased workload, and therefore, reducing the difficulty of the execution of countermeasures (i.e., simplifying them), will enhance their effects.

Seven different items were presented in our experiment, including one probe and six irrelevant items (following Hu et al., 2013). Out of the six irrelevant items, we chose to have two items for "oddballs," instead of only one, in order to raise the possibility that either one of them would evoke large enough P300 responses to defeat the test when using the P vs. Imax measure. In previous CTP studies on countermeasures, participants were instructed to execute different specific covert responses (typically: mentally say a meaningful personal name, e.g., close relative's given name) to each irrelevant that was targeted for countermeasures. Thus, to target a small, two-item group of irrelevant items, the original countermeasures could be used in the following way: silently articulating one specific name whenever one of the two small-group irrelevant item appears, and another specific name whenever the other small-group irrelevant item appears, while omitting covert responses whenever any of the other items appear (exactly as in the following studies: Hu et al., 2012; Labkovsky and Rosenfeld, 2012; Rosenfeld and Labkovsky, 2010).

In our version of countermeasures, participants were instructed to give concealed responses to all items: one silent word when either of the two irrelevant items that belonged to the "small group" appeared, and another silent word when any of the other items appeared, including the probe. This creates a simple, but continually active second task that may divert attention from the probe, which becomes, to some extent, simply the part of the larger group of irrelevant items. The recalling of the same word for these items of the larger group can also be described as giving them a common attribute, and thus making them overlapping. Consequently, the more the presented stimuli overlap with each other in their attributes, the smaller the P300 amplitude differences will be (Azizian et al., 2006; Marchand et al., 2013).

As the least detectable countermeasure, the silent, mental articulation of words was introduced in the study of Rosenfeld and Labkovsky (2010), and used in subsequent studies. These words were the first or last names of the participants, and, in some occasions, the first or last names of close relatives. Latency measures were not reported, but figures are provided on which it is consistently observable that, on the group level, P300 responses to probes and countered items peaked during the same time interval (Hu et al., 2012, Fig. 2; Labkovsky and Rosenfeld, 2012, Fig. 5; Rosenfeld and Labkovsky, 2010, Fig. 3; Winograd and Rosenfeld, 2011, Fig. 2). This shows that the P300 appears as a reaction to the item, and not to the subsequently recalled silent word, and it is therefore very likely that the effect on the P300 is caused not by the meaningfulness of the silent words, but by the meaningfulness of the countered items, i.e., that they are recognized as requiring a specific answer (Rosenfeld et al., 2013, p. 7, drew similar conclusions from some suggestive preliminary empirical evidence, citing the abstract of a yet unpublished study: Winograd and Rosenfeld, 2012). Possible emotional significance of names may have arousing effect, but higher arousal would simply lead to generally larger P300 responses throughout the task, including those to the probe (Duncan et al., 2009; Polich, 2007).

For the purpose of our above described countermeasure task, we asked our participants to choose any two simple, short (maximum two syllable) neutral words, that are easily distinguishable from each other, and which they would be comfortable to be repeating for 20–25 min, i.e., for the duration of the test.

1.4. Study outline

Four groups were measured: one "innocent" Control group, one "simple guilty" (SG) group with no instructions on countermeasures, and two other groups that were instructed to use countermeasures. One used our new countermeasures (New-CM group), and the other used the original countermeasures (Old-CM group) so that we could replicate previous findings, while also directly comparing the two countermeasure methods. In both groups, the countermeasure use involved choosing a small, two-item group of irrelevants; this set of items will be called the I-2item, while the remaining larger, four-item group of irrelevants will be called the I-4item. While the New-CM group used the above described countermeasures, the Old-CM group used countermeasures against the I-2item as described in previous studies, i.e., they were instructed to counter one of the irrelevant items by silently articulating the name of one of their parents, and another one with the name of the other parent (see in Methods; Section 2.2. Procedure).

To summarize our hypotheses: we expected a successful replication of the high accuracy rates with the P vs. Iall measure in all groups, but a significant drop in the detection rates with the P vs. Imax measure in the two CM groups (but not in the SG group) – and we expected this drop to be more pronounced with the enhanced countermeasures in the New-CM group.

2. Methods

2.1. Participants

Sixty-six participants were recruited through advertisements proposing to try our "EEG lie detection test," and offering a cafeteria voucher of Ft500 (approx. €1.60) in case they managed to defeat the test. Six participants were excluded: four due to excessive amount of artifacts in the EEG recording (over 50% of the trials had to be rejected), another due to extremely low accuracy in the decision task (correct responses to the catch trials of the "11111" item: 8.6%, correct responses in the cases of the four other strings of numbers: 91.8%), and one due to an extremely low rate of correct responses in the main task (66.9% correct). The remaining participants consisted of 14 individuals in the Control group (age = 24.5 ± 3.98 years, in the format of MEAN ± SD, as also in the rest of this paper; 5 males), 15 in the SG group (age = 21.5 ± 2.42 years; 6 males), 15 in the New-CM group (age = 22.9 ± 3.29 years; 7 males), and 16 in the Old-CM group (age = 23.0 ± 3.48 years; 6 males). All participants provided signed, informed consent, and, at the end of the experiment, they all received a cafeteria voucher regardless of the results of the examination.

2.2. Procedure

In this CIT, we used participants' family names as probes – except for the Control group, in which none of the presented items was relevant to the participants. For this group, we refer to the "Probe" item as the irrelevant item that was, unbeknownst to the participant, randomly assigned with the same EEG event marker as the Probe (own family

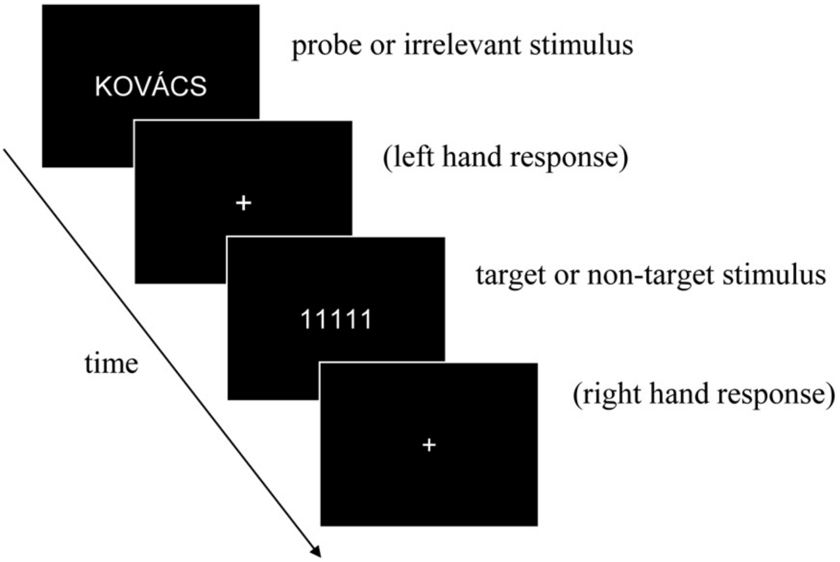

Fig. 1. Example of a trial in the CTP CIT task in this study. The probe and irrelevant stimuli were the given participant's own family name and other, unfamiliar family names. All these stimuli always required the same response with the left hand (pushing one randomly chosen key out of five, with the corresponding finger). This was followed by a target stimulus (11111) or a non-target stimulus (22222, 33333, 44444, and 55555). In response to a target stimulus, a key had to be pushed by the right middle finger, while in case of a non-target stimulus, another key had to be pushed by the right index finger.

name) of the other three experimental groups. At the beginning of each experiment, participants were shown a list of twenty Hungarian family names, and were asked to indicate if any of these names were particularly meaningful (e.g., name of a close relative or friend) or otherwise appeared to them markedly unique compared to the other names on the list. Irrelevant items were selected from among the family names that were not indicated by the given participant as salient.

The Old-CM group participants were instructed to use the given names of their parents to counter two out of the six irrelevant items (always recalling their father's name when one specific item appeared, and their mother's name when another specific item appeared). The New-CM group participants were asked to choose any two words that were short (one or two syllable) and easy to distinguish from each other; for example up/down or dog/cat. One of these words was recalled when either of two specific irrelevant item appeared, and the other was recalled when any of the other items (including the probe) appeared. We encouraged participants in this group to concentrate not on the item, but on which word-category the item belongs to. In both CM groups, the words used for countermeasure had to be silently said at the same time or after the response key was pushed. None of the participants had any information on the eventual irrelevant items in the task until the task began, and countermeasure using participants had to choose the small group of two specific irrelevant items during the beginning of the task.

The E-Prime software (Psychology Software Tools, Inc., Sharpsburg, USA) was used to present stimuli and record behavioral responses. Stimuli were presented in a 100 cm distance from the eye of the participant, on a 20 in. LCD screen. All presented characters were white on the black background, with a height subtending a visual angle of approximately 0.57°.

Each trial began with a 100 ms baseline period for the recording of prestimulus brain activity. The probe or irrelevant item (for the main task) was then presented on the center of the screen for 300 ms. Following an inter-stimulus interval that randomly varied between 1400–1700 ms, one of the number strings (for the secondary task) was presented for 300 ms. The next trial began after another randomly varying interval of 2100–2400 ms. During all intervals between stimuli, a fixation cross was presented on the center of the screen. For a schematic depiction, see Fig. 1.

In the full task, there were 350 trials in total, consisting of the probe (participant's family name) and six irrelevant items (other family names), each repeated 50 times, for a total of 50 probe and 300 irrelevant items presented in random order, followed by any of the number strings with equal probability (thus 10 times the 35 variations of the pairing of 7 names and 5 number strings).

Before the full task, countermeasure using participants were given a practice task that ran exactly the same way as the full task (1 probe and 6 irrelevant items presented in random order), except that participants assigned to the New-CM group were presented given names (with always the same names - “Ferenc” for males and “Ilona” for females - pointed out as probes[2]), while those in the Old-CM group were presented month names (with “January” as probe). Both CM groups were instructed to silently articulate the very same words (New-CM: two freely chosen neutral words; Old-CM: parents' given names) upon stimulus onset. This practice task had no time limit, participants could finish whenever they felt ready; this usually took 80–100 trials. As a next step, participants in all four experimental groups completed another practice task using a fixed set of city names (1 probe and 6 irrelevant items with “Budapest” as probe in all cases) that lasted for only 35 trials (5 repetitions for each stimulus). Here, participants were asked to imagine themselves in a real “lie detection” situation in which they would try to conceal the recognition of the probes. However, participants in the Control and SG groups were not instructed to use any specific countermeasures.

We removed most of the keys from a standard computer keyboard so that a group of five keys remained on the left side (to be conveniently pushed by each corresponding finger of the left hand) and two next to each other remained on the right side (to be pushed by the right index and right middle fingers). At each appearance of a probe or irrelevant item, a key had to be pushed with the left hand, each time randomly chosen from the five keys; participants were instructed not to follow any specific sequence (see Rosenfeld and Labkovsky, 2010). During the recording, we monitored responses to verify that participants were indeed randomly choosing keys, and not pushing the same key repeatedly or following a sequence (Rosenfeld and Labkovsky, 2010). In the secondary task, one of the keys on the right side had to be pushed

[2] These names were always different from the participants' own given names.

with the right hand at the appearance of a number string; one with the middle finger for the "11111" string, and the other one with the index finger when any of the other number strings appeared ("22222", "33333", "44444" or "55555").

Participants were told that from time to time during the task there would be a pause and they would be asked to repeat out loud the last family name that appeared, and more than two incorrect answers would mean failing the test. They were actually asked nine times during the experiment at random time points. None of the participants had more than two incorrect answers.

All participants in the three "guilty" groups were explicitly encouraged to try to conceal their names and to defeat the test. Participants completed the whole experiment (including preparations, practice and full tasks, and debriefing) in 50–60 min from their arrival.

2.3. Electrophysiological recordings and data processing

A BioSemi ActiveTwo Amplifier (BioSemi B. V., Amsterdam, Netherlands) was used with 32 Ag/AgCl electrodes recording EEG from standard scalp sites (Fp1, Fp2, AF3, AF4, F7, F3, Fz, F4, F8, FC5, FC1, FC2, FC6, T7, C3, Cz, C4, T8, CP5, CP1, CP2, CP6, P7, P3, Pz, P4, P8, PO3, PO4, O1, Oz, O2) and two additional electrodes placed above the left and right mastoids. The data was collected at a sampling rate of 1024 Hz, without any frequency filters. The recording reference and the ground electrodes (Common Mode Sense and Driven Right Leg electrodes in the ActiveTwo System; Metting van Rijn et al., 1990) were placed in close proximity to the Cz position.

Electrophysiological data were processed with the EEGLAB toolbox (Delorme and Makeig, 2004) for Matlab (MathWorks Inc., Natick, USA). After changing the sampling rate to 512 Hz (with Biosemi Decimator 86), the data was high-pass and low-pass filtered using Hamming-windowed sinc FIR filters with 0.3 Hz and 30 Hz cutoff frequencies, respectively (Widmann et al., 2015; Widmann and Schröger, 2012). Epochs starting at 100 ms before stimulus onset, and ending at 1300 ms after stimulus onset, were extracted, with baseline correction based on the whole epoch length. The entire recording was visually inspected for the removal of epochs with prominent artifacts such as baseline fluctuations or muscular activity. Ocular artifacts were removed with independent component analysis (ICA, Hyvärinen and Oja, 2000) implemented in EEGLAB. This method separated independent subcomponents of the EEG, among which those associated with eye movements were identified on the basis of visual inspection of their single-trial activations and scalp topography, and rejected. After applying a new baseline correction (from − 100 ms to 0 ms), the recording was again visually inspected to reject epochs with smaller artifacts. The mean and standard deviation of the remaining epochs for each stimulus (i.e., for each presented name for each participant) was 40.57 ± 6.34. Finally, the data was filtered again by applying a Hamming-windowed sinc FIR low-pass filter with 6 Hz cutoff frequency (Rosenfeld et al., 2008; Soskins et al., 2001), and the EEG was re-referenced to linked mastoids. For all statistical analyses, the P300 was measured at Pz only.

2.4. P300 measure and individual bootstrap analysis

For individual classification using P300 waves, a certain bootstrapping method has been used in all CTP studies, which compares the responses to the probe item with the responses to irrelevant items (see also: Farwell and Donchin, 1991; Wasserman and Bockenholt, 1989). This method uses a peak-to-peak measure (Rosenfeld, 2011; Rosenfeld et al., 2008; Soskins et al., 2001): in our case, an algorithm searched, on the averaged epoch of a certain stimulus type (as described below), for the maximum average 100 ms segment between 500 and 800 ms, and then, between the midpoint of this segment and 1300 ms, searched again for a minimum average 100 ms segment. The choice of the search window was based on visually inspecting the grand average of all participants, verifying that the P300 peak fell within the specified window (Keil et al., 2014; also cited by Rosenfeld et al., 2015b). The resulting value is the amplitude value of the peak-to-peak P300, which will be referred to as P300pp in the rest of this paper.

The procedure of the bootstrapping analysis for the P vs. Iall measure was the following. First, single trials were chosen randomly, with replacement, from all probe single trials (i.e., trials in which the probe item had been presented), and averaged into one epoch, from which a P300pp was calculated. The number of these chosen values was equal to the number of available probe trials in case of the given individual's results (i.e., the number of artifact-free epochs out of the original 50 recorded during the experiment). Second, a same number of single trials were again chosen randomly, with replacement, from all irrelevant single trials (i.e., trials in which one of the irrelevant items had been presented), and averaged into one epoch, from which another P300pp was calculated. Third, the P300pp obtained from the *probe* trials was compared to the P300pp obtained from *irrelevant* trials, in order to determine whether the former is greater than the latter (with a difference greater than zero). These three steps were repeated 1000 times, with results possibly varying according to the random choices with replacement. The end result of this procedure is a number between 0–1000, indicating the number of occasions in which the P300pp values of the probe trials were determined to be greater in comparison to those of the irrelevant trials.

The procedure for the P vs. Imax measure is exactly the same as the one for the P vs. Iall measure, except that the responses to the probe item were eventually compared to only one irrelevant item, the one which had evoked the largest P300pp, as measured with the bootstrap analysis. This individually varying largest irrelevant is called the "Imax." In order to select this Imax, an algorithm separately compared each of the six irrelevant items to the probe, which again resulted in a number between 0–1000, indicating the number of occasions in which the P300pp values of the probe were determined to be greater than those of the given irrelevant. The Imax was then selected from among all these six irrelevant items, to be the one in whose case this number was the smallest – and this smallest number is the result of the P vs. Imax measure for the given individual.

2.5. Group level comparisons

The distribution of behavioral data (mean item detection accuracy and reaction times) was entered into repeated-measures analyses of variance (ANOVA) with Stimulus Type (main task: probe vs. Iall; probe vs. I-2item vs. I-4item; secondary task: target vs. non-target) as within-subject factor and Group (I, SG, New-CM, Old-CM) as between-subject factor. The comparison between probe, I-2item and I-4item response latencies was necessary to show that any effects between probe vs. irrelevant items in the New-CM and/or Old-CM groups are due to the use of countermeasures.

Simple P300pp amplitudes – i.e., P300pp calculated from all single trials of the given stimulus type – were analyzed in three steps. First, a repeated-measures ANOVA was used to assess probe vs. Iall effects (Type as within-subject factor) between experimental groups. Then, a probe vs. Imax vs. Iremaining ANOVA was used to investigate the efficacy of the Imax measure in all four groups. Finally, with the probe vs. I-2item vs. I-4item statistical comparison between the two CM groups we aimed to show that the effect of countermeasures was more prominent in the New-CM group than in participants using the Old-CM technique.

Results of the individual bootstrap analysis (probe vs. Iall; probe vs. Imax) were used to classify participants as "innocent" or "guilty". One may set a cutoff rate, for example at 90% (Rosenfeld et al., 2013). In that case, when the P vs. Iall or P vs. Imax result for the individual is a number larger than 900 (i.e., the P300pp values of the probe trials were determined to be greater in >900 out of the 1000 calculations), then the participant is classified as guilty. For illustration, we report classification at several possible cutoffs (at 90%, 70%, and 50%), showing true negative rates (ratio of correctly identified innocent participants) in the case of the Control group and true positive rates (ratio of correctly identified guilty participants) in the cases of the SG and CM groups. However, for a more comprehensive assessment of classification accuracy, we calculated areas under the receiver operating characteristic curve (AUROC curve, or simply AUC – area under the curve; e.g., National Research Council, 2003, pp. 342–344). This method measures true positive and true negative rates at all possible cut-off points and gives an averaged value that can range from 0 to 1, where 0.5 means chance level classification, and 1 means flawless classification (i.e. all guilty and innocent classifications can be correctly made at a given cutoff point). The AUC was first calculated for the P vs. Iall results for each of the three guilty groups (SG, New-CM, Old-CM) paired with the P vs. Iall results of the Control group, and the resulting AUCs were compared using z tests (Hanley & McNeil, 1982). Finally, the same calculations and comparisons were made using the P vs. Imax results.

We used an alpha level of 0.05 for all statistical tests except for the bootstrapping measure. For each ANOVA with significant Group × Type interactions ($p < 0.05$), simple effects were tested using t-tests with Bonferroni correction. For violations of sphericity, Greenhouse-Geisser corrected p values and the relevant epsilon (ε) correction are reported. In order to demonstrate the magnitude of the observed effects, partial eta-squared (ηp^2) values are also shown.

3. Results

3.1. Behavioral measures

Accuracies and mean reaction times for the main and secondary tasks for all stimulus types and each experimental group are shown in Table 1.

In the main task, where all stimuli required the same response, mistakes and omitted responses were very rare in all conditions (see Table 1), and no statistically significant main effects or interactions were found ($p > 0.2$). In the secondary task, however, the participants' accuracies were significantly worse for target stimuli "11111" than for other strings (83 ± 2.2% vs. 98 ± 0.2%; F(1,56) = 51.7, $p < 0.001$, $\eta p^2 = 0.48$). This effect was not influenced by Group.

In the main task, the participants' reaction times were significantly slower for probe stimuli than for all the irrelevants (Iall) (546 ± 17 ms vs. 525 ± 18 ms; F(1,56) = 21.6, $p < 0.001$, $\eta p^2 = 0.28$), but this effect varied with Group significantly (F(3,56) = 5.6, $p = 0.002$, $\eta p^2 = 0.23$). Bonferroni-corrected tests of simple effects revealed that the probe vs. Iall comparison was significant for the SG ($p < 0.001$) and Old-CM groups ($p = 0.001$) only. The main effect of Group was not significant. In the two CM groups, the I-2item and I-4item stimuli were also compared with each other, and the probe. The significant Stimulus Type effect (F(2,58) = 12.1, ε = 0.754, $p < 0.001$, $\eta p2 = 0.29$) indicated that response times for probe (525 ± 23 ms) and I-2item (532 ± 27 ms) stimuli were substantially longer than those obtained for I-4item stimuli (488 ± 21 ms), with the Bonferroni post hoc tests being significant for both the P vs. I-4item ($p < 0.001$) and I-2item vs. I-4item ($p = 0.002$) comparisons. Again, the main effect of Group and its interaction with Stimulus Type (P vs. I-2item vs. I-4item) were not significant ($p > 0.2$).

Regarding the secondary task, responses to target stimuli were slower than to other stimuli (609 ± 12 ms vs. 539 ± 15 ms; F(1,56) = 88.8, $p < 0.001$, $\eta p^2 = 0.61$) and this effect was not influenced by Group. Interestingly, we have also found a significant Group main effect (F(3,56) = 3.2, $p = 0.035$, $\eta p^2 = 0.15$), but none of the post hoc comparisons reached significance level ($p > 0.072$).

Table 1
Means (M) and standard deviations (SD) for accuracies and reaction times (RT) to specific types of items, by each of the four groups.

	Groups							
	Control		SG		New-CM		Old-CM	
	M	SD	M	SD	M	SD	M	SD
Accuracies (%)								
Main task								
Probe	98.43	2.50	99.33	1.63	98.67	1.80	99.00	1.46
Iall	99.33	0.82	98.98	0.65	99.27	0.59	98.96	1.10
I-2item	–	–	–	–	99.27	0.70	98.63	1.67
I-4item	–	–	–	–	99.27	0.82	99.13	1.10
Secondary task								
Target	87.45	14.82	83.24	15.2	79.14	20.61	83.57	15.34
Non-target	98.85	1.14	97.69	2.23	97.62	2.01	98.39	0.96
RT (ms)								
Main task								
Probe	541	155	592	141	533	138	518	121
Iall	547	172	550	140	518	140	488	119
I-2item	–	–	–	–	542	154	524	157
I-4item	–	–	–	–	507	135	471	104
Secondary task								
Target	647	134	648	88	587	102	557	72
Non-target	595	163	577	108	509	127	479	88

Note. Main task: accuracies and RTs during the main task with family names. Probe – participant's own name; Iall – all names except the participant's own; I-2item – the two names that belonged to the smaller group of two items targeted by articulating the same words in the New-CM group (participants using the new countermeasures), and by articulating two different words in the Old-CM group (participants using the original countermeasures); I-4item – the four names that belonged to the larger group of four items that were targeted by articulating another word in the New-CM group, and simply omitted in the Old-CM group. There were no such groups of countered items (I-2item or I-4item) in the SG group (simple guilty participants) or in the Control group (innocent participants). Secondary task: accuracies and RTs during the secondary task with number strings. Target – the catch trials of "11111" strings that required response with the middle finger; Non-target – the rest of the number strings that required response with the index finger.

3.2. Electrophysiological measures

Event-related potentials obtained for probe, Iall and Imax stimuli for all four experimental groups are shown in Fig. 2, whereas means and standard deviations for P300pp amplitudes evoked by probe, Iall, I-2item and I-4item items are shown in Fig. 3.

3.2.1. P300pp amplitudes

As expected, Probe stimuli evoked significantly larger P300pp amplitudes than irrelevant (Iall) items (main effect of Stimulus Type: F(1,56) = 167.9, $p < 0.001$, $\eta p^2 = 0.75$; Fig. 3). Furthermore, there was a significant Stimulus Type × Group interaction (F(3,56) = 20.5, $p < 0.001$, $\eta p^2 = 0.52$). Bonferroni-corrected tests of simple effects indicated that the P300 evoked by Probe items was significantly smaller in the Control group than for SG participants ($p = 0.010$), and a trend was observed for the Control vs. Old-CM comparison ($p = 0.053$; and $p > 0.1$ for the rest of the comparisons between any two of the four groups). The P300 measured for Iall items did not differ between groups ($p > 0.999$). When performing post hoc Probe vs. Iall comparisons for each group separately, highly significant differences were found for the SG, New-CM and Old-CM groups ($p < 0.001$), while amplitudes were comparable in the Control group ($p = 0.728$). The main effect of Group was not significant.

When comparing the P300pp amplitudes evoked by Probe items, Imax items (irrelevant items with the largest P300pp), and Iremaining items (the means of the other five irrelevant items), the main effect of Stimulus Type was significant (F(2112) = 85.9, ε = 0.850, $p < 0.001$, $\eta p^2 = 0.61$): the largest P300pp means were for Probe, smaller for Imax, and smallest for Iremaining (for all comparisons, Bonferroni-corrected tests gave $p < 0.002$). The significant Stimulus Type × Group

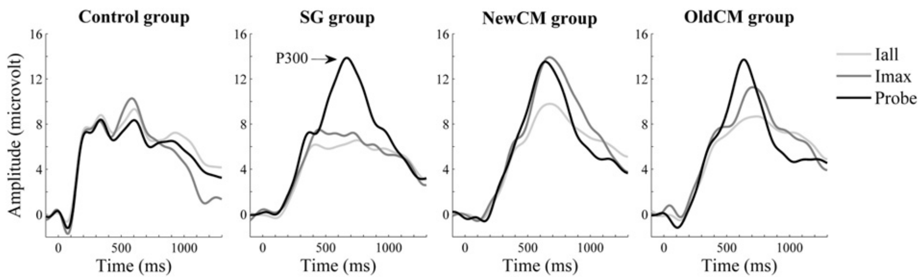

Fig. 2. Grand average event-related brain potential waveforms registered on the parietal electrode Pz, as evoked by the following stimuli: Probe (own family name), Imax (the one irrelevant family name that evoked the largest P300pp), and Iall (all irrelevant family names); within each of the four experimental groups: Control (innocent), SG (simple guilty), New-CM (participants using the new countermeasures), Old-CM (participants using the original countermeasures). Please note that on the group level, in the New-CM group, the P300pp evoked by the Imax item is even slightly larger than that evoked by the probe.

interaction (F(6112) = 14.2, $p < 0.001$, $\eta p^2 = 0.43$) indicated that the P300pp amplitudes were significantly larger for Probes than for Imax items in the SG ($p < 0.001$) and Old-CM (p < 0.001) groups, but were not found to be significantly different in the New-CM group ($p > 0.999$), and were significantly larger for Imax than for Probe items in the Control group ($p = 0.008$) (Fig. 2). Furthermore, the P300pp amplitudes were significantly larger for Probes than for Iremaining stimuli in the SG, Old-CM and New-CM groups, but not in the Control group (Bonferroni-corrected tests of simple effects for guilty groups: $p < 0.001$; for the Control group: $p = 0.645$). P300pp amplitudes were significantly larger for Imax than for Iremaining in the Control, New-CM and Old-CM groups ($p < 0.001$), while only a tendency was observed in for SG participants ($p = 0.067$). Again, the main effect of Group was not significant.

In order to test if increased P300pp amplitudes in the CM groups were indeed caused by countermeasure strategies, i.e., due to increased waveforms for the I-2item (the smaller group of countered irrelevant items), a second repeated-measures ANOVA was performed with Probe, I-2item, and I-4item stimuli as levels of Stimulus Type, and New-CM and Old-CM groups as levels of Group. The significant Stimulus Type × Group interaction (F(2,58) = 3.5, $p = 0.037$, $\eta p^2 = 0.11$) was indicative of robust P vs. I-2item amplitude differences in the Old-CM group only (Bonferroni-corrected tests of simple effects: $p < 0.001$), while similar amplitude reductions between I-2item vs. I-4item stimuli were observed in both CM groups (New-CM: $p = 0.033$, Old-CM: $p < 0.001$; Fig. 3). Finally, to see whether the difference between I-2item and I-4item differs in magnitude between New-CM and Old-CM groups, the ANOVA was rerun with the Probe omitted. The Stimulus Type (I-2item, I-4item) × Group interaction was not significant (F(1,29) = 1.5, $p = 0.227$, $\eta p^2 = 0.05$).

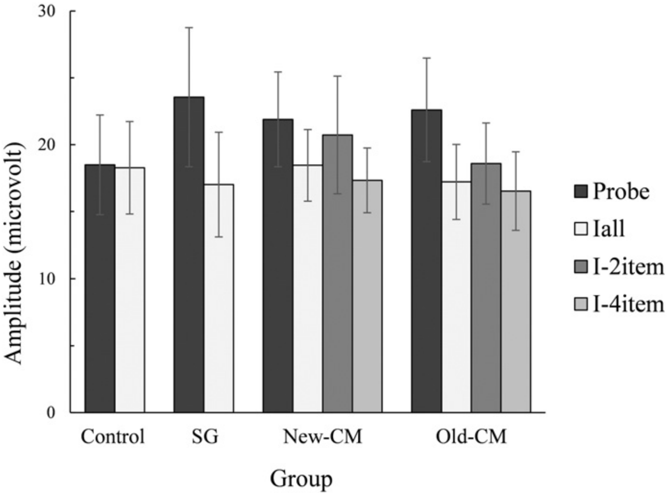

Fig. 3. Means and standard deviations of peak-to-peak P300 amplitudes registered on the parietal electrode Pz, for the following stimuli: Probe (own name), Iall (all irrelevant names), I-2item (the two names that belonged to the smaller group of two items targeted by articulating the same words in the New-CM group, and by articulating two different words in the Old-CM group), and I-4item (the four names that belonged to the larger group of four items that were targeted by articulating another word in the New-CM group, and simply omitted in the Old-CM group); within each of the four experimental groups: Control (innocent), SG (simple guilty), New-CM (participants using the new countermeasures), Old-CM (participants using the original countermeasures).

3.2.2. Individual classification based on the bootstrap analysis

Participants were classified guilty or innocent based on the results of the P vs. Iall and P vs. Imax measures – as described in Methods (Sections 2.4. and 2.5). Correct detection rates using cutoffs at 90%, 70%, and 50% are shown in Table 2, along with AUCs for each group, for P vs. Iall and for P vs. Imax, for which ROC curves are also shown in Fig. 4. In the case of classification using P vs. Iall measures (SG: AUC = 0.976, CI: 0.930–1; New-CM: AUC = 0.943, CI: 0.858–1; Old-CM: AUC = 0.929, CI: 0.831–1), no significant differences were found between the AUCs of any two of the three guilty groups ($p > 0.3$). In

Table 2
P vs. Iall and P vs. Imax bootstrap results for each participant.

Subject	P vs. Iall				P vs. Imax			
	Control	SG	NewCM	OldCM	Control	SG	NewCM	OldCM
1	964	1000	999	999	636	984	995	989
2	569	768	971	940	12	307	722[a]	291[a]
3	574	995	998	1000	183	983	952	994[a]
4	244	986	991	1000	11	865	727[a]	892
5	604	1000	1000	998	300	986	345[a]	897[a]
6	311	998	1000	992	50	972	996	415[a]
7	325	902	997	1000	6	578	78[a]	389[a]
8	780	1000	906	1000	230	997	272[a]	919[a]
9	699	1000	1000	975	332	990	969[a]	690
10	729	992	889	962	197	924	2[a]	347[a]
11	808	996	617	984	178	865	92[a]	492[a]
12	301	997	936	475	63	870	887	205[a]
13	793	995	905	999	98	942	710	982
14	679	1000	1000	1000	424	1000	515[a]	1000
15		1000	842	997		1000	103[a]	939[a]
16				748				259[a]
Mean	**599**	**975**	**936**	**942**	**194**	**884**	**557**	**669**
TR-9	0.93	0.93	0.80	0.88	1	0.67	0.27	0.38
TR-7	0.64	1	0.93	0.94	1	0.87	0.53	0.50
TR-5	0.29	1	1	0.94	0.93	0.93	0.60	0.56
AUC	–	**0.98**	**0.94**	**0.93**	–	**0.98**	**0.78**	**0.91**

Note. Group averages of the bootstrapping results are given below each corresponding column, in boldface (and rounded to whole numbers). TR: true negative rates (in the case of the Control group) and true positive rates (in the cases of the SG and CM groups) of correct individual classifications (number of correctly classified participants/number of participants) based on the P vs. Iall or P vs. Imax measures, with possible cutoffs, for illustration, at 900 (TR-9), 700 (TR-7), and 500 (TR-5) – where numbers below/above the given cutoff mean innocent/guilty classifications, respectively, for the given participant. AUC: areas under the curve for the three guilty (SG, Old-CM, New-CM) groups for the two measures (P vs. Iall and P vs. Imax), where each AUC uses the Control group's results of the same measure to calculate classification efficiency.

[a] One of the two countermeasure-target irrelevant items was the Imax.

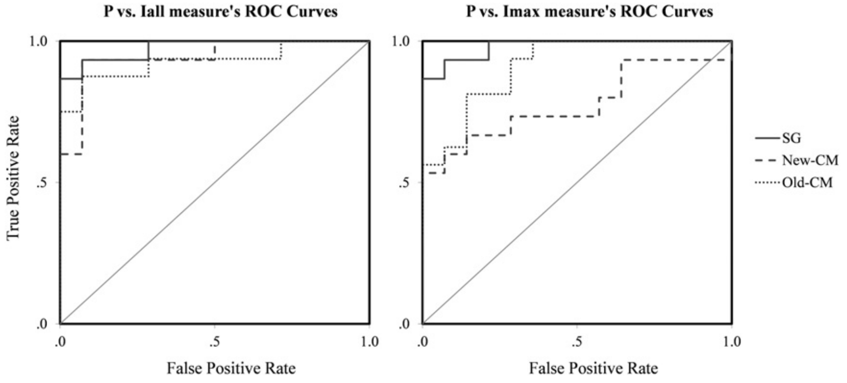

Fig. 4. ROC curves showing the true positives rates of the three guilty (SG, Old-CM, New-CM) groups in function of the false positive rates of the Control group (for the results of the P vs. Iall on the left, and for the results of the P vs. Imax on the right).

the case of P vs. Imax measures (SG: AUC = 0.981, CI: 0.943–1; New-CM: AUC = 0.776, CI: 0.598–0.954 for; Old-CM: AUC = 0.911, CI: 0.811–1), the AUC of the SG group was significantly larger than the AUC of the New-CM group (z = 2.21, p = 0.027) – meaning that the results of SG group's guilty participants, compared with the results of New-CM group's guilty participants, were significantly more distinct from the results of the Control group's innocent participants. No significant differences were found between the AUCs of the SG and the Old-CM groups (z = 1.29, p = 0.199) or between the AUCs of the New-CM and the Old-CM groups (z = −1.29, p = 0.196).[3]

4. Discussion

4.1. Effects of "small group" countermeasures on the P vs. Imax measure

The main purpose of our study was to show that the CTP version of the P300-based CIT, which has repeatedly been claimed to be highly resistant (or even immune) to countermeasures (Rosenfeld, 2011; Rosenfeld et al., 2013, 2008), can in fact be severely vulnerable to certain countermeasures, when using the P vs. Imax measure, i.e., when the probe (the critical information, e.g., crime detail) is compared to the Imax (the one irrelevant information which has evoked the largest P300pp responses). An effective countermeasure can be accomplished by covert responses to a small group of irrelevant items, and a different response, or no response at all, to all other items. This makes the items of the small group subjectively unique compared to the others, thereby evoking prominent P300pp waves, which can approximate or even overcome those evoked by the probe, reducing detection rates when using the P vs. Imax measure (Fig. 2).

To show this, for one, we have instructed the participants in the Old-CM group to use countermeasures that were used in previous studies, but whose effect, when used only on a smaller group of irrelevant items, has not been tested on the P vs. Imax measure, until now. For another, we have also introduced a slightly modified new version of these countermeasures, which was used by the participants in the New-CM group. In this latter case, all items were divided into a smaller and a larger group, and all items required a covert response according to group membership. The smaller group included two irrelevant items, targeted to be "oddballs," and the larger group included all the five other items (the probe and the four other irrelevant items). Furthermore, silent articulation of simple, easily distinguishable words (instead of personal names) were used for covert responses, in order to simplify the countermeasure task. The effects of this new countermeasure did prove to be somewhat more effective than the original one, and consequently helped us provide more convincing proof for the vulnerability of the P vs. Imax measure.

Our main results indeed show that participants in both CM groups have used the countermeasures against the P vs. Imax measure with success, generally achieving to be more difficult to distinguish from innocent participants, than those guilty participants who were not instructed to use countermeasures (AUC = 0.85 for CM participants vs. AUC = 0.98 for simple guilty participants) – though this difference was especially pronounced for those who used the new countermeasures (AUC = 0.78 for New-CM participants). Importantly, we also found, same as previous studies (Hu et al., 2012; Labkovsky and Rosenfeld, 2012; Rosenfeld et al., 2008; Rosenfeld and Labkovsky, 2010; Winograd and Rosenfeld, 2011), that the P vs. Iall measure (probe compared to all irrelevant items) provided high detection rates not only in the SG group (AUC = 0.98), but also in both CM groups (AUC = 0.93 using the original, and AUC = 0.94 using the new countermeasures), with no significant differences between the three groups. This also makes our study the first to show that the results of the P vs. Iall and the P vs. Imax measures, which have been shown to provide very similar accuracies in all previous experiments (Meixner et al., 2009; Meixner and Rosenfeld, 2014, 2011; Rosenfeld et al., 2008; Rosenfeld and Labkovsky, 2010), can in fact differ greatly.

The P vs. Imax measure was in fact never shown to provide substantially higher accuracies than the P vs. Iall measure. Consequently, we could suggest that the P vs. Imax measure, as a basis for guilty/innocent classifications, should simply be avoided in the future. On the other hand, our countermeasures increase false-negative rates for the P vs. Imax measure, but do not alter the fact that this measure provides high specificity (i.e., less false positive classifications; see Section 1.2; and, regarding our concurring results, see TR rows in Table 2 or the ROC curves in Fig. 4). Thus, the limitation arises when the probe P300pp is not found to be significantly larger than the Imax P300pp: in this case, the examinee may have used countermeasures, leading to a false negative classification. However, in practice, the P vs. Imax measure may still be useful to support the reliability of a positive finding: if the probe P300pp is not only significantly larger than the P300pp for the rest of the irrelevants, but also significantly larger than the Imax P300pp, than the guilty classification can be seen as more reliable.

We have noted in the Introduction (Section 1.2) that the CIT may also be used in cases when the probe is unknown. In this case, a suspect would be shown several items which are suspected to contain a relevant information (e.g., the possible locations of an upcoming terrorist attack), and the information that evokes the largest P300pp would be determined to be the presumed probe, whose P300pp is subsequently compared to those of the others (Meixner and Rosenfeld, 2011; Rosenfeld, 2011). In our study, we first looked for the Imax among the

[3] In order to demonstrate the overall effect of countermeasures, we also performed additional AUC calculations for the CM groups merged into one CM group that can be defined simply as "participants instructed to use countermeasures" (resulting in AUC = 0.935, CI: 0.865–0.1 using P vs. Iall; and AUC = 0.846, CI: 0.733–0.959 using P vs. Imax). In the case of the P vs. Imax measure, the AUC of the SG group was significantly larger than the AUC of this merged CM group (z = 2.23, p = 0.026), i.e., the P vs. Imax results of this merged CM group were more distinct from those of the Control group. Again, no significant differences were found in the case of the P vs. Iall measure (p > 0.3).

irrelevant items, and then compared it to the probe. Logically, in each case when this Imax proved to evoke a P300pp larger than that of the probe, the Imax was in fact also proven to have evoked the largest P300pp among all items (thus also including the actual probe), and therefore would have been, in an unknown-probe scenario, incorrectly selected as the presumed probe. According to our results, out of the 31 countermeasure user participants, 13 (42%) would have succeeded in making us select a wrong item (an irrelevant) to be the presumed probe (see Table 2, where the P vs. Imax iterations below 500 signify that, in the bootstrap analysis, the probe P300pp was determined to be smaller than the Imax; which was the case in 6 cases out of 15 in the New-CM group, and in 7 cases out of 16 in Old-CM, although only in 1 case out of 15 in the SG group; as also shown in the TR-5 row in Table 2). In such a case, we may run another analysis, comparing the presumed probe to irrelevant items (either by P vs. Iall or P vs. Imax measure), and accordingly classify the examinee as innocent or guilty – but, in the case of a guilty classification, we would have, unfortunately, no way of knowing whether the presumed probe (the item, which evoked the largest P300pp responses), is the actual probe, or a countered irrelevant. That makes the unknown-probe scenario, as described in the article of Meixner and Rosenfeld (2011), highly vulnerable to countermeasures. The solution for this problem awaits further studies on this matter.

Our results corroborate previous findings by Rosenfeld and his colleagues, including, most importantly, very high detection accuracy, and therefore we conclude that the replication of the CTP protocol was successful. However, further independent studies would be needed for a thorough validation of this method. In particular, it should be noted that we have used personal items (family names) as probes in our study, which, in the case of P300-based studies, generally leads to higher detection accuracies when compared to crime details (e.g., a weapon used in a recent crime; Meijer et al., 2014). Therefore, future studies should also assess the validity of the method when using crime details, e.g., in a mock-crime, preferably in field settings.

Finally, we want to point out a methodological issue that could have introduced a minor confound to our results. Namely, different stimuli were presented in the first practice task to the New-CM and Old-CM groups (given names and month names, respectively), which might have facilitated countermeasure application for New-CM participants in the main task, since it also relied on names. Although we argue that this is very unlikely because (1) the stimulus sets seen by the New-CM group were completely different (practice task: given names, main task: family names) and (2) the two tasks were separated by a second practice task that used city names in all experimental groups, we acknowledge that it would have been better to use the same stimulus set (e.g., month names) for training in both CM groups.

4.2. Summary

In this study, we have shown that the P vs. Imax measure of the CTP method (Rosenfeld et al., 2013, 2008) can be defeated by covertly creating a small group of “oddball” items among the presented irrelevant items, thereby making their P300pp responses statistically indistinguishable from those of the probe item. We have also shown that countermeasures can be further enhanced for this specific reason. Although these countermeasures strongly reduced detection rates when using the P vs. Imax measure, our results corroborated previous studies in that the P vs. Iall measure provided high detection rates in all groups, and thereby proved to be resistant to both the original and the modified countermeasures.

Acknowledgements

Author G. Cs. is supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences and by the Hungarian National Brain Research Program (Grant no. KTIA 13 NAP-A-II/20). The cafeteria vouchers were provided by the Student's Union of the Faculty of Arts at the University of Szeged. We thank Júlia Simon for her assistance during pilot studies.

References

Azizian, A., Freitas, A.L., Parvaz, M.A., Squires, N.K., 2006. Beware misleading cues: perceptual similarity modulates the N2/P3 complex. Psychophysiology 43, 253–260. http://dx.doi.org/10.1111/j.1469-8986.2006.00409.x.

Ben-Shakhar, G., 2012. Current research and potential applications of the concealed information test: an overview. Front. Psychol. 3. http://dx.doi.org/10.3389/fpsyg.2012.00342.

Bond, C.F., DePaulo, B.M., 2008. Individual differences in judging deception: accuracy and bias. Psychol. Bull. 134, 477–492. http://dx.doi.org/10.1037/0033-2909.134.4.477.

Bond, C.F., DePaulo, B.M., 2006. Accuracy of deception judgments. Personal. Soc. Psychol. Rev. Off. J. Soc. Personal. Soc. Psychol. Inc 10, 214–234. http://dx.doi.org/10.1207/s15327957pspr1003_2.

Delorme, A., Makeig, S., 2004. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J. Neurosci. Methods 134, 9–21. http://dx.doi.org/10.1016/j.jneumeth.2003.10.009.

Donchin, E., 1981. Surprise!? Surprise? Psychophysiology 18, 493–513. http://dx.doi.org/10.1111/j.1469-8986.1981.tb01815.x.

Donchin, E., Coles, M.G.H., 1988. Is the P300 component a manifestation of context updating? Behav. Brain Sci. 11, 357. http://dx.doi.org/10.1017/S0140525X00058027.

Duncan, C.C., Barry, R.J., Connolly, J.F., Fischer, C., Michie, P.T., Näätänen, R., Polich, J., Reinvang, I., Van Petten, C., 2009. Event-related potentials in clinical research: guidelines for eliciting, recording, and quantifying mismatch negativity, P300, and N400. Clin. Neurophysiol. 120, 1883–1908. http://dx.doi.org/10.1016/j.clinph.2009.07.045.

Farwell, L.A., Donchin, E., 1991. The truth will out: interrogative polygraphy (“lie detection”) with event-related brain potentials. Psychophysiology 28, 531–547. http://dx.doi.org/10.1111/j.1469-8986.1991.tb01990.x.

Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143 (1), 29–36. http://dx.doi.org/10.1148/radiology.143.1.7063747.

Hartwig, M., Bond, C.F., 2011. Why do lie-catchers fail? A lens model meta-analysis of human lie judgments. Psychol. Bull. 137, 643–659. http://dx.doi.org/10.1037/a0023589.

Hu, X., Hegeman, D., Landry, E., Rosenfeld, J.P., 2012. Increasing the number of irrelevant stimuli increases ability to detect countermeasures to the P300-based complex trial protocol for concealed information detection: countermeasures in P300 memory detection. Psychophysiology 49, 85–95. http://dx.doi.org/10.1111/j.1469-8986.2011.01286.x.

Hu, X., Pornpattananangkul, N., Rosenfeld, J.P., 2013. N200 and P300 as orthogonal and integrable indicators of distinct awareness and recognition processes in memory detection: N200–P300 in memory detection. Psychophysiology 50, 454–464. http://dx.doi.org/10.1111/psyp.12018.

Hyvärinen, A., Oja, E., 2000. Independent component analysis: algorithms and applications. Neural Netw. Off. J. Int. Neural Netw. Soc. 13, 411–430.

Katayama, J.'i., Polich, J., 1999. Auditory and visual P300 topography from a 3 stimulus paradigm. Clin. Neurophysiol. 110, 463–468. http://dx.doi.org/10.1016/S1388-2457(98)00035-2.

Keil, A., Debener, S., Gratton, G., Junghöfer, M., Kappenman, E.S., Luck, S.J., Luu, P., Miller, G.A., Yee, C.M., 2014. Committee report: publication guidelines and recommendations for studies using electroencephalography and magnetoencephalography: guidelines for EEG and MEG. Psychophysiology 51, 1–21. http://dx.doi.org/10.1111/psyp.12147.

Kraut, R., 1980. Humans as lie detectors. J. Commun. 30, 209–218. http://dx.doi.org/10.1111/j.1460-2466.1980.tb02030.x.

Labkovsky, E., Rosenfeld, J.P., 2012. The P300-based, complex trial protocol for concealed information detection resists any number of sequential countermeasures against up to five irrelevant stimuli. Appl. Psychophysiol. Biofeedback 37, 1–10. http://dx.doi.org/10.1007/s10484-011-9171-0.

Lykken, D.T., 1959. The GSR in the detection of guilt. J. Appl. Psychol. 43, 385–388. http://dx.doi.org/10.1037/h0046060.

Marchand, Y., Inglis-Assaff, P.C., Lefebvre, C.D., 2013. Impact of stimulus similarity between the probe and the irrelevant items during a card-playing deception detection task: the “irrelevants” are not irrelevant. J. Clin. Exp. Neuropsychol. 35, 686–701. http://dx.doi.org/10.1080/13803395.2013.819837.

Meijer, E.H., Selle, N.K., Elber, L., Ben-Shakhar, G., 2014. Memory detection with the concealed information test: a meta analysis of skin conductance, respiration, heart rate, and P300 data: CIT meta-analysis of SCR, respiration, HR, and P300. Psychophysiology 51, 879–904. http://dx.doi.org/10.1111/psyp.12239.

Meissner, C.A., Kassin, S.M., 2002. “He's guilty!”: investigator bias in judgments of truth and deception. Law Hum. Behav. 26, 469–480.

Meixner, J.B., Haynes, A., Winograd, M.R., Brown, J., Rosenfeld, J.P., 2009. Assigned versus random, countermeasure-like responses in the P300 based complex trial protocol for detection of deception: task demand effects. Appl. Psychophysiol. Biofeedback 34, 209–220. http://dx.doi.org/10.1007/s10484-009-9091-4.

Meixner, J.B., Rosenfeld, J.P., 2014. Detecting knowledge of incidentally acquired, real-world memories using a P300-based concealed-information test. Psychol. Sci. 25, 1994–2005. http://dx.doi.org/10.1177/0956797614547278.

Meixner, J.B., Rosenfeld, J.P., 2011. A mock terrorism application of the P300-based concealed information test: mock terrorism concealed information test. Psychophysiology 48, 149–154. http://dx.doi.org/10.1111/j.1469-8986.2010.01050.x.

Mertens, R., Allen, J.J.B., 2008. The role of psychophysiology in forensic assessments: deception detection, ERPs, and virtual reality mock crime scenarios. Psychophysiology 45, 286–298. http://dx.doi.org/10.1111/j.1469-8986.2007.00615.x.

Metting van Rijn, A.C., Peper, A., Grimbergen, C.A., 1990. High-quality recording of bioelectric events. Part 1. Interference reduction, theory and practice. Med. Biol. Eng. Comput. 28, 389–397.

National Research Council, 2003. Polygraph and Lie Detection. The National Academies Press, Washington, D.C.

O'Connell, R.G., Dockree, P.M., Kelly, S.P., 2012. A supramodal accumulation-to-bound signal that determines perceptual decisions in humans. Nat. Neurosci. 15, 1729–1735. http://dx.doi.org/10.1038/nn.3248.

Polich, J., 2007. Updating P300: an integrative theory of P3a and P3b. Clin. Neurophysiol. 118, 2128–2148. http://dx.doi.org/10.1016/j.clinph.2007.04.019.

Rosenfeld, J.P., 2011. P300 in detecting concealed information. In: Verschuere, B., Ben-Shakhar, G., Meijer, E. (Eds.), Memory Detection: Theory and Application of the Concealed Information Test. Cambridge University Press, Cambridge.

Rosenfeld, J.P., Cantwell, B., Nasman, V.T., Wojdac, V., Ivanov, S., Mazzeri, L., 1988. A modified, event-related potential-based guilty knowledge test. Int. J. Neurosci. 42, 157–161.

Rosenfeld, J.P., Hu, X., Labkovsky, E., Meixner, J., Winograd, M.R., 2013. Review of recent studies and issues regarding the P300-based complex trial protocol for detection of concealed information. Int. J. Psychophysiol. 90, 118–134. http://dx.doi.org/10.1016/j.ijpsycho.2013.08.012.

Rosenfeld, J.P., Labkovsky, E., 2010. New P300-based protocol to detect concealed information: resistance to mental countermeasures against only half the irrelevant stimuli and a possible ERP indicator of countermeasures: P300 CTP resists two mental countermeasures. Psychophysiology no–no. http://dx.doi.org/10.1111/j.1469-8986.2010.01024.x.

Rosenfeld, J.P., Labkovsky, E., Winograd, M.R., Lui, M.A., Vandenboom, C., Chedid, E., 2008. The complex trial protocol (CTP): a new, countermeasure-resistant, accurate, P300-based method for detection of concealed information. Psychophysiology 45, 906–919. http://dx.doi.org/10.1111/j.1469-8986.2008.00708.x.

Rosenfeld, J.P., Soskins, M., Bosh, G., Ryan, A., 2004. Simple, effective countermeasures to P300-based tests of detection of concealed information. Psychophysiology 41, 205–219. http://dx.doi.org/10.1111/j.1469-8986.2004.00158.x.

Rosenfeld, J.P., Ward, A., Frigo, V., Drapekin, J., Labkovsky, E., 2015a. Evidence suggesting superiority of visual (verbal) vs. auditory test presentation modality in the P300-based, complex trial protocol for concealed autobiographical memory detection. Int. J. Psychophysiol. 96, 16–22. http://dx.doi.org/10.1016/j.ijpsycho.2015.02.026.

Rosenfeld, J.P., Ward, A., Thai, M., Labkovsky, E., 2015b. Superiority of pictorial versus verbal presentation and initial exposure in the P300-based, complex trial protocol for concealed memory detection. Appl. Psychophysiol. Biofeedback 40, 61–73. http://dx.doi.org/10.1007/s10484-015-9275-z.

Soskins, M., Rosenfeld, J.P., Niendam, T., 2001. Peak-to-peak measurement of P300 recorded at 0.3 Hz high pass filter settings in intraindividual diagnosis: complex vs. simple paradigms. Int. J. Psychophysiol. Off. J. Int. Organ. Psychophysiol. 40, 173–180.

Twomey, D.M., Murphy, P.R., Kelly, S.P., O'Connell, R.G., 2015. The classic P300 encodes a build-to-threshold decision variable. Eur. J. Neurosci. 42, 1636–1643. http://dx.doi.org/10.1111/ejn.12936.

Verleger, R., Baur, N., Metzner, M.F., Śmigasiewicz, K., 2014. The hard oddball: effects of difficult response selection on stimulus-related P3 and on response-related negative potentials: oddball-P3 and S–R mapping. Psychophysiology 51, 1089–1100. http://dx.doi.org/10.1111/psyp.12262.

Verleger, R., Hamann, L.M., Asanowicz, D., Śmigasiewicz, K., 2015. Testing the S–R link hypothesis of P3b: the oddball effect on S1-evoked P3 gets reduced by increased task relevance of S2. Biol. Psychol. 108, 25–35. http://dx.doi.org/10.1016/j.biopsycho.2015.02.010.

Verschuere, B., Ben-Shakhar, G., Meijer, E., 2011. Memory Detection: Theory and Application of the Concealed Information Test. Cambridge University Press, Cambridge.

Verschuere, B., Meijer, E.H., 2014. What's on your mind?: recent advances in memory detection using the concealed information test. Eur. Psychol. 19, 162–171. http://dx.doi.org/10.1027/1016-9040/a000194.

Wasserman, S., Bockenholt, U., 1989. Bootstrapping: applications to psychophysiology. Psychophysiology 26, 208–221.

Widmann, A., Schröger, E., 2012. Filter effects and filter artifacts in the analysis of electrophysiological data. Front. Psychol. 3. http://dx.doi.org/10.3389/fpsyg.2012.00233.

Widmann, A., Schröger, E., Maess, B., 2015. Digital filter design for electrophysiological data — a practical approach. J. Neurosci. Methods 250, 34–46. http://dx.doi.org/10.1016/j.jneumeth.2014.08.002.

Winograd, M.R., Rosenfeld, J.P., 2014. The impact of prior knowledge from participant instructions in a mock crime P300 concealed information test. Int. J. Psychophysiol. 94, 473–481. http://dx.doi.org/10.1016/j.ijpsycho.2014.08.002.

Winograd, M.R., Rosenfeld, J.P., 2012. Countermeasure mechanisms in the complex trial protocol. Int. J. Psychophysiol. 85, 305. http://dx.doi.org/10.1016/j.ijpsycho.2012.06.046.

Winograd, M.R., Rosenfeld, J.P., 2011. Mock crime application of the Complex Trial Protocol (CTP) P300-based concealed information test: mock crime CTP P300 concealed information test. Psychophysiology 48, 155–161. http://dx.doi.org/10.1111/j.1469-8986.2010.01054.x.