UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Sprengholz, P., Schreckenbach, F., Giesen, C. G., Koranyi, N., & Rothermund, K. (2023). Guilty on the Go: Uncovering Concealed Information by Assessing Response Preparation Processes in a Go-Nogo-Paradigm. *Collabra: Psychology*, *9*(1). https://doi.org/10.1525/collabra.77819

Cognitive Psychology

Guilty on the Go: Uncovering Concealed Information by Assessing Response Preparation Processes in a Go-Nogo-Paradigm

Philipp Sprengholz[1,2][a], Franziska Schreckenbach[1], Carina G. Giesen[3], Nicolas Koranyi[1], Klaus Rothermund[1]

[1] University of Jena, Jena, Germany, [2] University of Bamberg, Bamberg, Germany , [3] HMU Health and Medical University, Erfurt, Germany

Keywords: concealed information test, guilty knowledge, lie detection, faking, response preparation

https://doi.org/10.1525/collabra.77819

Collabra: Psychology

Vol. 9, Issue 1, 2023

The Reaction Time-Based Concealed Information Test (RT-CIT) was designed to detect familiarity with crime-related information. However, RT-CIT results can be manipulated by preparing innocent-looking responses for these probes. We developed a new paradigm allowing us to assess such response preparation processes. In each trial of the task, a crime-related prime question was presented, followed by a test item which was either a publicly known item, an irrelevant item, or a probe. The test item could either match the question in terms of content or not, and a key had to be pressed if the answer was not publicly known (Go-Nogo task). In two experiments (total $N = 138$), we found evidence for both familiarity (slower reactions toward crime-related than unrelated information) and response preparation effects (less errors for probes matching the prime question) in guilty participants, indicating that the new paradigm can help to overcome problems of strategic response preparation.

Introduction

As humans cannot reliably detect deceptive communication (Feeley & Young, 1998), crime investigators must generally rely on indirect evidence such as fingerprints, DNA, and polygraph results to convict suspects. As one of the most prominent polygraph techniques, the Concealed Information Test (CIT) is designed to uncover guilty knowledge (Lykken, 1959, 1960). Suspects are presented with critical crime-related items such as objects or photos from the crime scene that have not previously been leaked, along with irrelevant items unrelated to the crime in question. While innocent parties cannot reliably distinguish between critical and irrelevant stimuli, a guilty person can, and recognition of crime-related details can be measured using physiological indicators such as skin conductance or event-related potentials (Meijer et al., 2014).

Research has shown that paradigms comparing response latencies provide a valid index of deception too (Verschuere et al., 2010). The Reaction Time Based CIT (RT-CIT) typically employs an oddball task, in which participants use a keyboard to respond to a sequence of briefly presented stimuli (Seymour et al., 2000; Seymour & Kerlin, 2008; Visu-Petra et al., 2012), which are of three types: probes, irrelevant items, and targets. *Probes* are critical details about the crime that could only be recognized by the culprit; among the non-critical information, *irrelevant items* have not been encountered before while *targets* are memorized prior to the oddball task. In each trial, participants are asked to react as quickly as possible by pressing a specified key if the presented stimulus is a target and an alternate key if it is not. For innocent participants, this is a binary task that depends on stimulus familiarity (Verschuere & De Houwer, 2011); presentation of a familiar target elicits one response while presentation of either a critical but unfamiliar probe or an irrelevant item triggers the other.

A purely familiarity-based classification would result in fast reactions for all three stimulus types (Yonelinas, 2002). However, as both targets and probes are familiar to guilty participants, basing the response on familiarity alone would lead to an incorrect response whenever a probe is presented. Monitoring responses for the probes should slow down reaction times and increase errors in classification of probes as compared to irrelevant items because the automatic familiarity-based response must be overridden by an explicit decision to conceal knowledge of the probe (Verschuere & De Houwer, 2011). This RT-CIT effect is similar in size to physiology-based CIT effects (Suchotzki et al., 2017).

As reaction time tasks require less expensive technology and training than recording physiological signals, the RT-CIT has been advocated by many researchers and practi-

[a] Correspondence concerning this article should be addressed to:
Philipp Sprengholz, University of Bamberg, Markusstr. 8a, 96045 Bamberg, Germany
E-Mail: philipp.sprengholz@uni-bamberg.de

tioners. In principle, however, reaction times are prone to manipulation strategies motivated by the intention to appear innocent. Guilty suspects can mimic the behavior of innocents using two alternative strategies. The first option is to slow down one's responses to irrelevant items. For example, Rosenfeld et al. (2004) showed that RTs could be increased by countermeasures such as pressing a finger or imagining a slap in the face when irrelevant items are presented. As a consequence, RTs for irrelevant items could no longer be separated from RTs for probes, and test accuracy dropped from 91% to 45%. However, Suchotzki et al. (2021) recently showed in a high-powered study that response deadlines help to prevent the use of this strategy. The second strategy is to improve the classification response to probes (in terms of RT and error) by preparing the required responses before probes are presented. According to relevant accounts of stimulus-response binding and retrieval (Hommel, 1998, 2004, 2013; see also Frings et al., 2020, for a theoretical update), when a response is executed in temporal contiguity to a stimulus, an episode comprising stimulus and response is stored in episodic memory. Transferring this account of S-R binding and retrieval to the domain of lying, Koranyi et al. (2015) showed that when a question is answered with a lie, knowledge about having lied to this question is automatically retrieved from memory when the question comes up again (Koranyi et al., 2015). A similar effect was found for suppressed or omitted information, resulting in automatic retrieval of former lies of omission (Schreckenbach et al., 2020).

Based on these findings, we assume that the retrieval of critical information and associated responses should also influence how participants respond during the RT-CIT. When a guilty suspect is asked during interrogation about a probe from the crime scene, they must override and correct their initial familiarity-based responses, replacing these with innocuous responses in order to appear innocent. Similarly, when preparing for an RT-CIT, guilty suspects who anticipate that critical items will be presented during the test may form "implementation intentions" (Gollwitzer & Sheeran, 2006), linking these stimuli to innocent-looking responses in order to conceal their familiarity with these items. These strategic preparations form stimulus-response episodes that connect the critical item to the response which is required in the task. These are subsequently retrieved from memory when the stimulus appears again, triggering automatic activation of the planned response (Martiny-Huenger et al., 2017). By facilitating responses to probes, this retrieval effect reduces differences in response latencies for probes and irrelevant items, which corrupts the rationale of the familiarity effect and diminishes RT-CIT sensitivity.

Herein, we present a novel paradigm that captures familiarity-based effects (as in the RT-CIT) but simultaneously reveals response preparation processes, explicit or implicit (automatic) response alterations that might be used to conceal spontaneous reactions to critical items. It follows that guilty knowledge can no longer be fully concealed using response strategies, as it will either show up in familiarity-based effects specific to probes or—if these effects are suppressed by response preparation processes—in automatic retrieval effects that are specific for probes too.

In the novel paradigm, a priming question about the crime is followed by a go/no-go task in which a key must be pressed to suppress probes and irrelevant items that have not been leaked or memorized while no key is to be pressed if a target appears. For instance, when investigating the theft of exams and manuscripts from a university server, suspects might be asked to memorize the fact that exams have been stolen. When the target word *exams* is presented in the go/no-go task, no action is required; when an item is presented that was not part of the memorizing instructions, a go key must be pressed. However, this go response applies both to irrelevant items such as *payslips* (which might have been stolen but were not), and to probes like *manuscripts*, which were stolen but were not part of the memorizing instructions. While this is an easy task for innocent participants who know nothing about the critical (stolen) items (e.g., *manuscripts*), a guilty participant would be expected to respond differently to probes and irrelevant items in a go trial: the higher familiarity of the probes means these should be mistaken as targets, which require a no-go response, so delaying go responses to probes. This effect structurally reflects a relevant S-R compatibility effect (see De Houwer, 2003; Verschuere & De Houwer, 2011) since novelty temporarily becomes part of the Go response, which has to be pressed in order to stop officially unknown information to be expressed in response to a question. The familiarity of the guilty knowledge probes conflicts with the to-be-executed response since concealing guilty knowledge requires participants to press the STOP key (i.e., declaring the item as not known or unfamiliar) for these probes despite them being familiar.

To conceal these familiarity-based delays, guilty suspects might develop a strategy that prepares a fast go response to the probes. If successful, this strategy will counteract familiarity-based slowing, diminishing or neutralizing differences between probes and irrelevant items. Importantly, such strategies can only be developed by guilty suspects, as only they can distinguish between probes and irrelevant items. Detection of such strategies is therefore another way of discovering guilty suspects and prevents them from eluding the paradigm. To detect the preparation of innocent-looking responses for probes, we introduced questions as primes in the go/no-go task. These questions refer to specific sets of probe items; for instance, the answers matching the question *What was stolen?* are the above-mentioned *exams, manuscripts,* and *payslips*. Each question pre-activates certain probes but pre-activates different answers for guilty and innocent persons. In the case of innocent suspects, only *exams* will be activated by the question *What was stolen?* For guilty suspects, however, the same question will activate both *exams* and *manuscripts*. If a guilty suspect has prepared a go response for the probe *manuscripts*, the question will also activate the go response they prepared for this item. To expose these prepared responses, go responses to probes after matching and non-matching prime questions are compared. As prepared go responses for probes are elicited by matching questions,

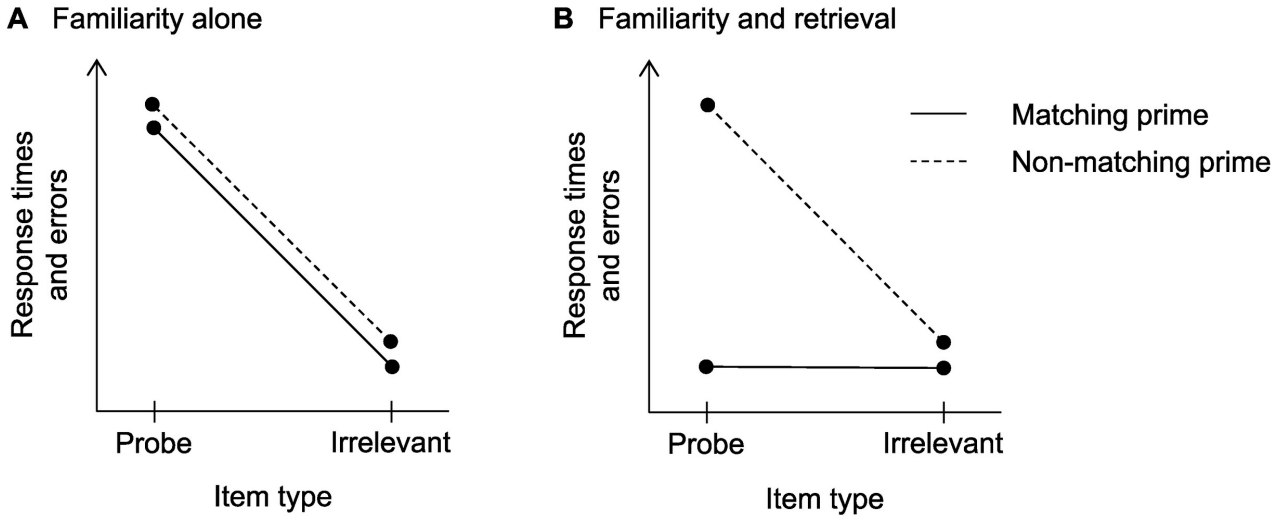

Figure 1. Hypothesized effects.

(A) For guilty suspects, the familiarity effect is reflected in increased response latencies and error rates for probes as compared to irrelevant items. (B) The retrieval effect is reflected in an interaction of prime question and probe type. Automatic activation of prepared go responses for probes by matching questions should reduce reaction times and error rates for probes after matching as compared to non-matching prime questions.

faster go responses for probes after matching than after non-matching prime questions should be observed. On the other hand, no such difference should obtain for irrelevant items for which no go responses have been prepared, as irrelevant items are unknown to both guilty and innocent individuals. Importantly, this difference in responding to probes and irrelevant items after matching questions should obtain only for guilty suspects; as only they can distinguish between probes and irrelevant items, only they can prepare go responses for probes.

On that basis, we formulated the following predictions regarding guilty suspects for the four possible combinations of probes and irrelevant items with matching and non-matching prime questions. (a) When responses are based on familiarity alone, responses to probes will be delayed regardless of whether the priming question is matching or non-matching, resulting in a main effect of probe type (Figure 1A). (b) When guilty suspects prepare go responses for probes, there will be an interaction effect; prepared responses will mute the familiarity effect, but this counteracting effect of prepared go responses will be stronger after matching questions (Figure 1B). By specifying these two indicators of concealed knowledge, the proposed paradigm should prove more informative in identifying guilty suspects than relying on the familiarity effect alone (as employed by the RT-CIT).

To validate the paradigm, we performed two online experiments employing a mock crime scenario. In the first experiment, we investigated only the response patterns of guilty participants. We predicted (a) a main effect for probes vs. irrelevant items because of the familiarity of the probes, and (b) an interaction of probe type and prime question because automatic retrieval of prepared go responses for probes after matching prime questions neutralizes the probe effect in the matching prime condition. In the second experiment, results were replicated, and a control group of innocent examinees was included to assess accuracy of classification.

Experiment 1

Methods

Participants and Design

The experiment employed a 3 (probe vs. target vs. irrelevant item) x 2 (matching vs. non-matching of priming question and stimulus) within-subjects design. Participants were recruited in January 2020 at a German university, using the psychology department's mailing list. In total, N = 58 undergraduates completed the experiment, exceeding the required minimum sample size (n = 54) to identify medium-sized effects (f = .25, α = .05, 1 – β = .80, assuming a default correlation of r = .05 between effects for probes and irrelevant items) as calculated using G*Power 3 (Faul et al., 2007). The participants (43 female, 15 male) ranged in age from 18 to 38 years (M = 21.57, SD = 3.93). All were offered partial course credit in exchange for their participation.

Procedure and Materials

The experiment comprised three stages. In the first stage, participants were instructed to commit a mock crime by hacking a server at the university and downloading some documents. To obtain the password and access the files, they had to run two tools (e.g., *Medusa* and *Grover*) in a console window. On completing this task, they discovered that the files came from two departments (e.g., *bioinformatics* and *philosophy*) and included two types of documents (e.g., *manuscripts* and *exams*). The tools, departments, and documents employed in a given case were randomly selected from six such options (see online supplement). For

instance, the documents might equally have been *payslips, contracts, certificates,* or *transcripts*.

In the second stage of the experiment, participants were told that they were under suspicion for the hack and should try to escape a digital investigator. The investigator first asked them to read a press release about the hack containing only half the details of what participants would have encountered when committing the crime (e.g., that a tool named *Grover* was used to hack the university server in order to steal *exams* from the department of *philosophy*). The items named in the press release were the targets, and the unnamed crime-relevant items were the probes. Information that did not form part of the crime scene or press release served as irrelevant items. After reading the press release, a chat window opened, and the digital investigator asked for general personal information (e.g., *How old are you?*), as well as three crime-related questions (e.g., *What has been stolen?*). The critical questions could be answered by providing information from the press release, so allowing participants to conceal guilty knowledge from the investigator.

In the final stage of the experiment, participants completed a go/no-go task (Figure 2). Each trial started with a fixation cross (400ms), followed by a priming question, which was always one of the three critical questions posed during the chat (1000ms). A possible answer was then presented as a probe stimulus; this was either a target from the press release (e.g., *exams*), a probe (e.g., *manuscripts*), or an irrelevant item (e.g., *payslips*). After 1500ms, the word automatically moved toward the investigator at the top of the screen (500ms), indicating the utterance as an answer to the previously shown question. Participants were instructed to utter nothing but target words; to cancel the utterance, they had to press the space key (go response) before the text arrived at the investigator. This allowed us to measure response times when withholding probes (to appear innocent) and irrelevant items (as instructed). When participants made a mistake (e.g., by pressing the key when a target word appeared or not pressing when a probe or irrelevant item was shown), a message from the investigator indicated that they were behaving suspiciously. A blank screen was shown for 1000ms between trials. To ensure correct encoding of the material, attention checks were inserted after some trials; the likelihood that a participant would be asked about the prime question or an answer shown in the current trial before the next trial started was 7.5%.

Trials were administered in two blocks. In Block 1 (96 trials), prime questions and answer stimuli always matched (e.g., *What was stolen?* was followed by *manuscripts*) in order to establish a strong expectation that questions would be followed by matching probes. The study's full 2 x 2 design was realized in Block 2 (168 trials); that is, answers matched the prime question in the majority of trials (144 trials or 86%) but were non-matching in the remaining 24 trials (e.g., *What was stolen?* was followed by *bioinformatics*). Responses (go vs. no-go) were counterbalanced for both types of question, with 50% no-go (targets), and 50% go responses (25% probes, 25% irrelevant items) appearing after each type. A list of all trial combinations can be found in the online supplement. The experiment was programmed and run using jsPsych (de Leeuw, 2015).

Results

The analysis of response times and error rates for probes and irrelevant items in Block 2 revealed that error rates in the go/no-go tasks (M = 2.5%) and attention checks (M = 3.5%) were below 25% for all participants, which failed to justify exclusion from further analyses. We discarded response times (RTs) of erroneous responses and those that were below 250ms or more than three interquartile ranges above the third quartile of an individual's RT distribution (far-outs; see Tukey, 1977), amounting in total to 1.7 % of all response times.

RTs and error rates were compared for matching and non-matching trials and for probes and irrelevant items, using two-way repeated measures ANOVAs (Table 1). A main effect of stimulus type emerged for both RT and error data; on average, responses to probes were slower and more error-prone than for irrelevant items. A significant interaction was observed for error rates only; responses to probes (M = 2.6%) were more error-prone than to irrelevant items (M = 0.0%) when presented after a non-matching question (t[57] = 2.88, p = .006) as compared to a matching question (probes: M = 0.6%, irrelevant items: M = 0.6%, t < 1). This confirms the hypothesized retrieval effect (Figure 3).

Importantly, the effects for error rates were not stable when removing the first non-matching trials for each item type and participant. In this case, both the main effect of item type and the interaction of item and prime failed to reach statistical significance (both p = .11, see online supplement).

Discussion

The results of the first experiment align with previous findings regarding the RT-CIT (Verschuere & De Houwer, 2011). The rejection of familiar probes in a go/no-go task was found to require additional time and produced more errors than for irrelevant items as a consequence of familiarity, which is similar to what is observed in standard RT-CIT tasks with binary responses (e.g., Verschuere & De Houwer, 2011). Most importantly, we also found evidence of strategic response preparation for probes. Comparing the effects of item type (probes vs. irrelevant items) for matching and non-matching prime questions, we found that error rates for probes were higher than for irrelevant items when presented after non-matching prime questions. However, this effect was eliminated for matching prime questions. The interaction effect suggests that automatic retrieval of prepared responses for probes was triggered by corresponding questions (cf. Koranyi et al., 2015; Schreckenbach et al., 2020).

Interestingly, the interaction effect of item and prime type on error rates depended on participants' reaction toward the first non-matching trials. When excluding these trials from the analyses, the effects turned insignificant. This finding might indicate that effects relating to the type

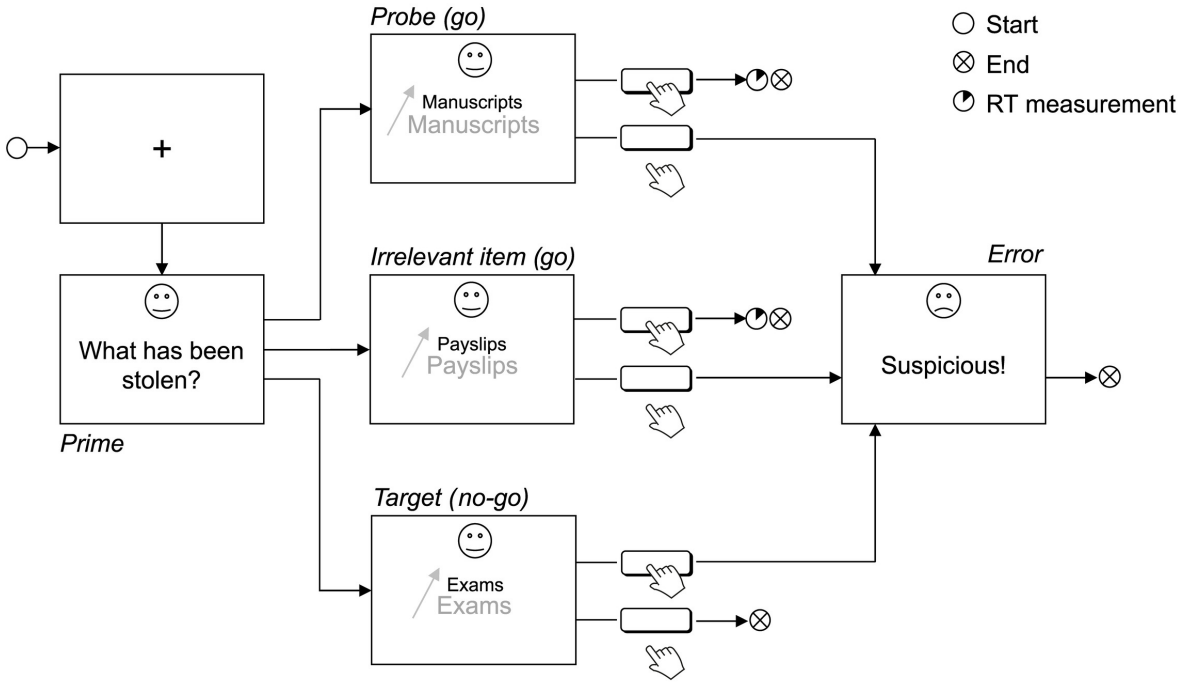

Figure 2. Trial structure of the go/no-go task.

Table 1. Experiment 1 ANOVA results

		RT			Error rate		
	df	*F*	*p*	η_G^2	*F*	*p*	η_G^2
Item type: probe vs. irrelevant item (familiarity effect)	57	**26.47**	< .001	.03	**8.69**	.005	.03
Prime type: Matching vs. non-matching question	57	3.74	.058	.00	2.98	.090	.01
Interaction of prime and probe type (retrieval effect)	57	0.95	.334	.00	**7.13**	.010	.03

Note: Results from two-way repeated measures ANOVAs. Bold values are statistically significant with $p < .05$.

of prime question are short-lived and can be overcome by repeated presentations. However, the large differences in error rates between first and later non-matching stimuli may also stem from the first block of trials where only matching stimuli were presented. As participants got used to the matching, the reactions toward the first non-matching stimuli may have been particularly revealing with regard to the response preparation processes. We therefore tested whether the problem also appears if the first matching-only block is removed in a second experiment. In this second experiment we also investigated whether the retrieval effect can be used to distinguish between guilty and innocent suspects. As the retrieval effect for probes indicates specific knowledge of those items, it should allow identification of guilty suspects—especially where familiarity effects are weak or non-existent.

Experiment 2

Unlike Experiment 1, in which every participant committed a mock crime, Experiment 2 included a control group of innocent participants who committed no crime. This enabled us to assess our paradigm's accuracy of classification by comparing the response patterns of guilty and innocent participants. The experiment was preregistered[1] and reused large parts of the material from Experiment 1.

1 The preregistration protocol can be found at https://aspredicted.org/d9g2w.pdf. Please note that we originally planned to analyze Inverse Efficiency Scores (IES) only, as has been done in previous studies using different paradigms (e.g., Koranyi et al., 2015; Schreckenbach et al., 2020). However, as analyzing reaction times and error rates separately provided more information about how the new paradigm is working, we decided to report these measures here. IES results align with error rate results and can be found in the online supplement.

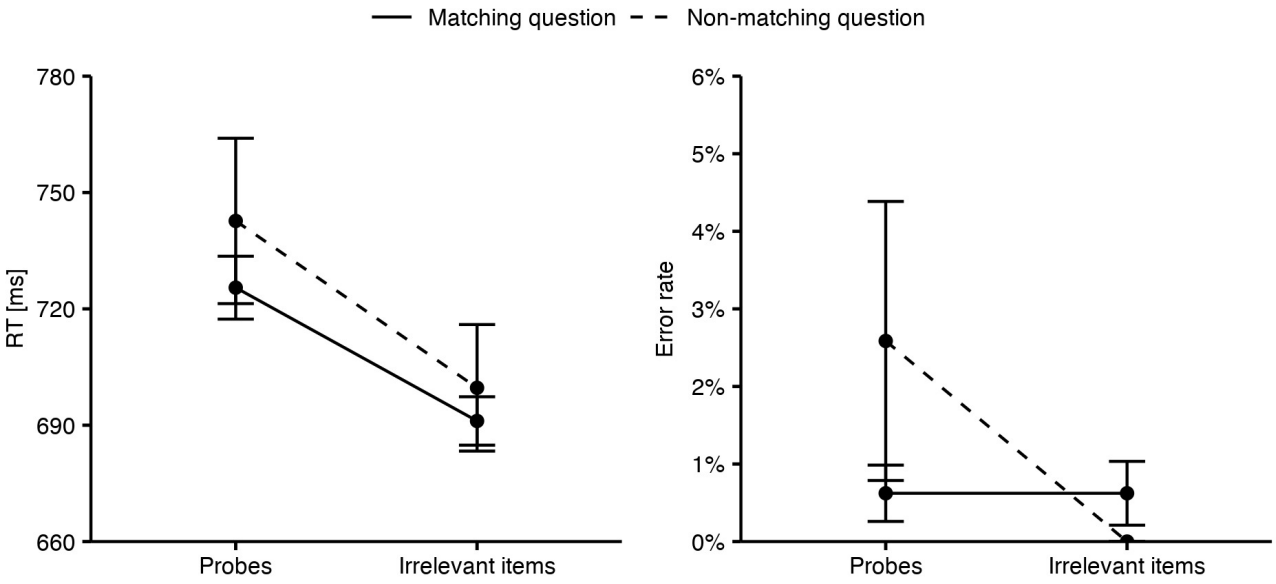

Figure 3. Experiment 1 pattern of means and 95% confidence intervals for RTs and error rates by item type (probes vs. irrelevant items) and prime question (matching vs. non-matching).

Methods

Participants and Design

The experiment employed a 2 (guilty vs. innocent group) x 3 (probe vs. target vs. irrelevant stimulus) x 2 (matching vs. non-matching of prime and stimulus) design. Group assignment was random, and the other two factors were varied within subjects. Participants were recruited in October 2020 at the same German university, using the psychology department's mailing list. In total, N = 80 undergraduates (40 per group) completed the experiment, exceeding the required minimum total sample size (n = 54) to identify medium-sized effects per group (f = .25, α = .05, 1 – β = .95, assuming a default correlation of r = .05 between effects for probes and irrelevant items) according to G*Power 3 (Faul et al., 2007). Participants ranged in age from 18 to 28 years (M = 20.93, SD = 2.77). Of these, 68 were female, 10 were male, and 2 did not indicate their gender. All were offered partial course credit in exchange for their participation.

Procedure and Materials

The procedure and materials were the same as in Experiment 1, with two important changes. First, participants who were randomly assigned to the innocent group did not engage in the mock crime but were instead asked to evaluate the usability of the university website. Second, to shorten the experiment, go/no-go trials from Block 1 were replaced by 20 (always matching) tutorial trials. The 168 trials of the original experimental block remained unchanged. Again, the experiment was programmed and run using jsPsych (de Leeuw, 2015).

Results

For all participants, error rates in the go tasks (M = 1.2%), no-go tasks (M = 7.2%) and attention checks (M = 4.3%) were below 25%. Based on the same criteria as in Experiment 1, 0.8% of all trials were regarded as RT outliers and were discarded from all analyses. RTs and error rates were submitted to 2 (stimulus type: probe vs. irrelevant item) x 2 (priming question: matching vs. non-matching) x 2 (guilty vs. innocent subjects) mixed models ANOVAs. These revealed a significant three-way interaction for error rates, $F(1,78) = 8.44$, $p = .005$, $\eta_G^2 = 0.02$, but not for response times, $F(1,78) = 2.29$, $p = .134$, $\eta_G^2 = 0.00$.

We also investigated effects separately for innocent and guilty participants, using two-way ANOVAs to compare RTs and error rates for matching and non-matching trials and between probes and irrelevant items. For innocent participants (Table 2, top panel), RTs and error rates did not differ significantly across conditions (all $F < 2.19$, $p > .147$). For guilty participants, however, response times and error rates were higher for probes, indicating a familiarity effect.[2] Most importantly, we found a significant interaction effect for error rates (Table 2, bottom panel), indicating that guilty participants committed more errors in probe trials (M = 3.1%) than in irrelevant item trials (M = 1.5%) if prime question

2 The effect for errors was significant only in a one-tailed test, but such a test corresponds to our directional hypothesis that was based on previous findings with the RT-CIT.

Table 2. Experiment 2 ANOVA results

		RT			Error rate		
	df	*F*	*p*	η_G^2	*F*	*p*	η_G^2
Innocent participants							
Item type: probe vs. irrelevant item (familiarity effect)	39	0.53	.469	.00	0.15	.701	.00
Prime type: Matching vs. non-matching question	39	2.19	.147	.00	3.80	.059	.02
Interaction of prime and probe type (retrieval effect)	39	0.27	.605	.00	0.16	.690	.00
Guilty participants							
Item type: probe vs. irrelevant item (familiarity effect)	39	**16.42**	< .001	.05	3.56	.067	.01
Prime type: Matching vs. non-matching question	39	**4.29**	.045	.01	3.11	.086	.02
Interaction of prime and probe type (retrieval effect)	39	2.51	0.12	.00	**15.87**	< .001	.04

Note: Results from two-way repeated measures ANOVAs. Bold values are statistically significant with $p < .05$.

and probe did not match ($t[158] = 1.67$, $p = 0.048$) while no such difference was observed after matching questions (probes: $M = 0.9\%$; irrelevant items: $M = 1.1\%$, $t < 1$) (see Figure 4). Importantly, the interaction effect remained significant when excluding the first non-matching trial for each item type and every participant from analyses (see online supplement), indicating the stability of the effect. In sum, the error patterns of guilty participants replicated the findings of Experiment 1, indicating both familiarity and retrieval effects, but no such effects were observed for innocent participants.

To investigate whether the new paradigm could differentiate between innocent and guilty subjects, individual familiarity and retrieval effects were calculated for RTs and error rates, and two logistic models were validated using the leave-one-out method. In the first model, group membership was predicted by familiarity effects for RTs and error rates (calculated as individual differences between standardized RTs/error rates for probes and irrelevant items), resulting in a classification accuracy of 56.3%. In the second model, the list of predictors was extended to include retrieval effects for RTs and error rates (calculated as individual differences between differences of RTs/error rates for non-matching probes and irrelevant items and differences of RTs/error rates for matching probes and irrelevant items). This increased classification performance to 66.3%. Compared with the first model (AIC = 108.14), the second model better fitted the data (AIC = 99.38). In short, classification was improved by considering both familiarity and retrieval effects.

Discussion

The second experiment replicated and extended the results from Experiment 1; again, we found evidence for the response preparation processes when participants were guilty. While response times were generally higher for probes than for irrelevant items, higher error rates were observed only when probes followed non-matching question primes. After matching prime questions, this effect was eliminated, indicating retrieval of prepared no-go responses for probes. As expected, no such effects could be found for innocent participants. As in Experiment 1, the response preparation effect of guilty participants relied on error rates only. Contrary to our prediction, no interaction effect could be found for response times. Future work should investigate possible explanations for this discrepancy. On one hand, it may result from a measurement problem. As both experiments were conducted online and response time measurement in web browsers can be assumed to show more latency and variance than in a lab environment, small effects may not be detected. Consequently, the experiments should be replicated in the lab. Alternatively, the pattern of results may stem from the instructions used in the experiment. While participants were instructed to make no errors, the go/no-go design stressed rapid responses by introducing a response deadline for the go responses. Therefore, participants may have focused on minimizing reaction times for all trials, resulting in a floor effect, and conceded making errors in the minority of non-matching trials. Similar shifts of systematic variance from RTs to errors are often obtained in speeded response tasks using a response deadline (e.g., Draine & Greenwald, 1998; Greenwald et al., 1996; Musch & Klauer, 2001; Wentura & Degner, 2010).

Looking beyond the findings of the previous experiment, we found that the classification of innocent and guilty participants improved significantly when not only familiarity effects but also retrieval effects were included as predictors in the regression. This confirms that the new paradigm's ability to simultaneously assess familiarity-based and retrieval-based effects for probes circumvents faking outcomes and improves discriminant validity.

General Discussion

We presented a new paradigm for detecting guilty knowledge that builds on the idea of the conventional RT-CIT. In line with previous research about the RT-CIT (Verschuere & De Houwer, 2011), crime-related probes led to a familiarity-based increase in response times and error

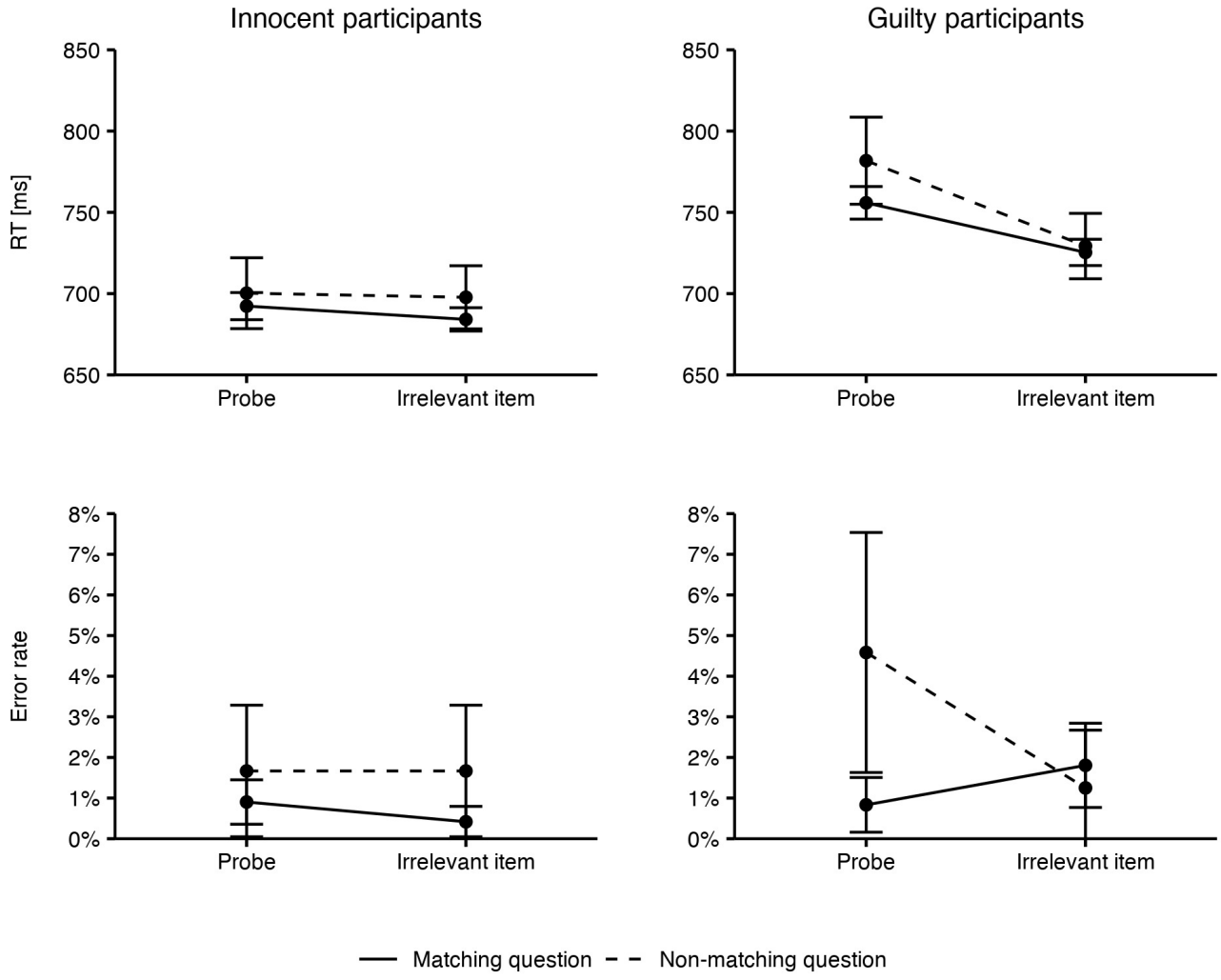

Figure 4. Experiment 2 pattern of means and 95% confidence intervals for RTs and error rates by probe type (probes vs. irrelevant items) and prime question (matching vs. non-matching), analyzed separately for guilty and innocent participants.

rates among guilty suspects as compared to irrelevant information. The new paradigm was also designed to detect a specific strategy to manipulate the outcome of the task by guilty participants: preparing innocent-looking responses to probes in order to conceal guilty knowledge. This strategy has previously been suggested to undermine the validity and reliability of the RT-CIT (Verschuere & Meijer, 2014). By including questions as prime stimuli that were presented immediately before the probes enabled us to elicit and measure retrieval processes for prepared responses for probes that were triggered by the respective prime questions (cf. Koranyi et al., 2015; Schreckenbach et al., 2020). Matching questions eliminated the familiarity effect for probes. By exposing response preparation for probes, this type of manipulative strategy can be detected. Faked responses to probes are in themselves indicative of guilty knowledge and can be used to identify guilty suspects. In line with these considerations, the results of our second experiment revealed that using retrieval-based effects to predict group membership improved the ability to discriminate between guilty and innocent suspects over and above familiarity-based effects. Importantly, participants in both experiments were not instructed on how the paradigm discriminates between guilty and innocent suspects, which is a prerequisite for the usual definition of faking behavior in the literature. While we think that even understanding the test would not help suspects to fake responses without being found out (as it should not be possible to counteract familiarity effects with prepared responses without simultaneously being identified), future studies should evaluate the paradigm under more realistic conditions and check whether the classification performance changes when participants have been informed about how the paradigm is working.

As the present study did not include any direct comparison of the standard RT-CIT and the new paradigm, we cannot make strong or absolute claims regarding their relative and absolute performance, which remains to be investigated in future research. However, as the new paradigm and the RT-CIT assess familiarity-based effects, the former may outperform the RT-CIT, as the combination of famil-

iarity-based and retrieval-based effects was shown to outperform predictions based on familiarity effects alone.

Previous research has shown that a multiple probes protocol for the RT-CIT is superior to variants of the task that use only a single probe (Verschuere et al., 2015). This advantage of the multiple probes protocol can be explained in terms of response preparation processes, which are much easier and more efficient if they refer to only a single critical item (e.g., preparing a "No" response to the question "Do you recognize this item?"). However, although more complex and probably less efficient, response preparation can influence results also in the multiple probes protocol of the RT-CIT, by preparing "No" responses for multiple different items. To uncover these strategies, we developed the new paradigm, which allows an assessment of familiarity effects (as in the RT-CIT), but simultaneously also allows for an identification of response preparation strategies. Our paradigm deviates from the RT-CIT in several relevant respects, most importantly by replacing the standard two-alternative forced choice task (do you recognize this item? - yes vs. no) with a go/nogo task. The introduction of varying prime questions also slowed down responding in the new task, which might have reduced the reliance on familiarity, and thus weakened the familiarity effect. In the future, it might thus be worthwhile to develop a task that combines the rationale of our new paradigm with the two-alternative forced choice task that has previously been used in the RT-CIT (measuring RTs also for targets as in the RT-CIT) in order to assesses response preparation by including matching and non-matching primes such as crime-relevant questions (as in the presented paradigm).

A major limitation of the new paradigm is that it not only needs multiple crime information categories (e.g., what has been stolen, from whom, using what tool) but also two crime relevant items per category (one to be used as the target, and one as the critical probe). Such information is not always available in real investigations, limiting the use of the new paradigm. Another limitation of our study is that we validated the paradigm only in the context of mock crimes. For research and validation purposes, this strategy has some advantages in relation to the investigation of real crimes, where a suspect's guilt or innocence is typically uncertain, and a specific variant of the task has to be designed for each crime scenario. Nevertheless, the ultimate aim is to provide information about guilty knowledge, and the novel paradigm will ultimately have to demonstrate that it can predict guilt under more realistic circumstances. Clearly, with classification accuracy rates below 70%, the novel paradigm's discrimination performance does not yet match the quality criteria for reliable categorization of individual suspects. However, classification rates may improve under more realistic conditions involving more extreme familiarity effects and strategic attempts to conceal those effects. As Kleinberg & Verschuere (2019) could show that finding the optimal statistical separation between lies and truths in a single dataset can produce overoptimistic accuracy rates, future research on the proposed paradigm should employ a cross-validation approach across multiple experiments. This also requires a closer look at how the paradigm performs under more realistic base rates. In Experiment 2, guilty and innocent participants were distributed evenly but in real crime investigations more innocent than guilty suspects may be tested. Therefore, future research should assess both sensitivity and specificity in scenarios with different base rates.

While more research is certainly needed to assess the performance and predictive validity of the novel paradigm under more realistic conditions, one promising aspect of the new procedure is that it facilitates detection of response preparation for critical items, which could be an efficient faking strategy for this kind of test. As well as protecting the test from being compromised, the ability to identify this strategy also provides another more indirect means of identifying guilty knowledge, which is required to devise such a faking strategy in the first place. To that extent, the promising findings reported here mark an important step toward the improvement and wider application of techniques for the reliable detection of concealed knowledge.

Ethical Declaration

Both experiments have been conducted in accordance with ethical standards and were approved by the Ethical Commission of the University of Jena (FSV 19/44). All participants provided informed consent prior to data collection.

Contributions

PS, FS, CG, NK, and KR designed the research; PS performed research; PS and KR planned and performed data analysis; PS wrote the initial draft, which was revised and approved by all authors.

Competing Interests

The authors declare that they have no conflicts of interest.

Data Accessibility Statement

Materials, data, and data analysis scripts are available at https://dx.doi.org/10.17605/OSF.IO/MP37Z

Submitted: January 31, 2023 PDT, Accepted: May 15, 2023 PDT

References

De Houwer, J. (2003). A structural analysis of indirect measures of attitudes. In J. Musch & K. C. Klauer (Eds.), *The psychology of evaluation: Affective processes in cognition and emotion* (pp. 219–244). Lawrence Erlbaum Associates Publishers.

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Draine, S. C., & Greenwald, A. G. (1998). Replicable unconscious semantic priming. *Journal of Experimental Psychology: General*, *127*(3), 286–303. https://doi.org/10.1037/0096-3445.127.3.286

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/bf03193146

Feeley, T. H., & Young, M. J. (1998). Humans as lie detectors: Some more second thoughts. *Communication Quarterly*, *46*(2), 109–126. https://doi.org/10.1080/01463379809370090

Frings, C., Hommel, B., Koch, I., Rothermund, K., Dignath, D., Giesen, C., Kiesel, A., Kunde, W., Mayr, S., Moeller, B., Möller, M., Pfister, R., & Philipp, A. (2020). Binding and Retrieval in Action Control (BRAC). *Trends in Cognitive Sciences*, *24*(5), 375–387. https://doi.org/10.1016/j.tics.2020.02.004

Gollwitzer, P. M., & Sheeran, P. (2006). Implementation Intentions and Goal Achievement: A Meta-analysis of Effects and Processes. *Advances in Experimental Social Psychology*, 69–119. https://doi.org/10.1016/s0065-2601(06)38002-1

Greenwald, A. G., Draine, S. C., & Abrams, R. L. (1996). Three Cognitive Markers of Unconscious Semantic Activation. *Science*, *273*(5282), 1699–1702. https://doi.org/10.1126/science.273.5282.1699

Hommel, B. (1998). Event Files: Evidence for Automatic Integration of Stimulus-Response Episodes. *Visual Cognition*, *5*(1–2), 183–216. https://doi.org/10.1080/713756773

Hommel, B. (2004). Event files: Feature binding in and across perception and action. *Trends in Cognitive Sciences*, *8*(11), 494–500. https://doi.org/10.1016/j.tics.2004.08.007

Hommel, B. (2013). Dancing in the dark: no role for consciousness in action control. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00380

Kleinberg, B., Arntz, A., & Verschuere, B. (2019). Being accurate about accuracy in verbal deception detection. *PLOS ONE*, *14*(8), e0220228. https://doi.org/10.1371/journal.pone.0220228

Koranyi, N., Schreckenbach, F., & Rothermund, K. (2015). The implicit cognition of lying: Knowledge about having lied to a question is retrieved automatically. *Social Cognition*, *33*(1), 67–84. https://doi.org/10.1521/soco.2015.33.1.67

Lykken, D. T. (1959). The GSR in the detection of guilt. *Journal of Applied Psychology*, *43*(6), 385–388. https://doi.org/10.1037/h0046060

Lykken, D. T. (1960). The validity of the guilty knowledge technique: The effects of faking. *Journal of Applied Psychology*, *44*(4), 258–262. https://doi.org/10.1037/h0044413

Martiny-Huenger, T., Martiny, S. E., Parks-Stamm, E. J., Pfeiffer, E., & Gollwitzer, P. M. (2017). From conscious thought to automatic action: A simulation account of action planning. *Journal of Experimental Psychology: General*, *146*(10), 1513–1525. https://doi.org/10.1037/xge0000344

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. *Psychophysiology*, *51*(9), 879–904. https://doi.org/10.1111/psyp.12239

Musch, J., & Klauer, K. C. (2001). Locational uncertainty moderates affective congruency effects in the evaluative decision task. *Cognition & Emotion*, *15*(2), 167–188. https://doi.org/10.1080/02699930126132

Rosenfeld, J. P., Soskins, M., Bosh, G., & Ryan, A. (2004). Simple, effective countermeasures to P300-based tests of detection of concealed information. *Psychophysiology*, *41*(2), 205–219. https://doi.org/10.1111/j.1469-8986.2004.00158.x

Schreckenbach, F., Sprengholz, P., Rothermund, K., & Koranyi, N. (2020). How to Remember Something You Didn't Say. *Experimental Psychology*, *67*(6), 364–372. https://doi.org/10.1027/1618-3169/a000504

Seymour, T. L., & Kerlin, J. R. (2008). Successful detection of verbal and visual concealed knowledge using an RT-based paradigm. *Applied Cognitive Psychology*, *22*(4), 475–490. https://doi.org/10.1002/acp.1375

Seymour, T. L., Seifert, C. M., Shafto, M. G., & Mosmann, A. L. (2000). Using response time measures to assess "guilty knowledge". *Journal of Applied Psychology*, *85*(1), 30–37. https://doi.org/10.1037/0021-9010.85.1.30

Suchotzki, K., Verschuere, B., & Gamer, M. (2021). How vulnerable is the reaction time concealed information test to faking? *Journal of Applied Research in Memory and Cognition*, *10*(2), 268–277. https://doi.org/10.1016/j.jarmac.2020.10.003

Suchotzki, K., Verschuere, B., Van Bockstaele, B., Ben-Shakhar, G., & Crombez, G. (2017). Lying takes time: A meta-analysis on reaction time measures of deception. *Psychological Bulletin*, *143*(4), 428–453. https://doi.org/10.1037/bul0000087

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.

Verschuere, B., Crombez, G., Degroote, T., & Rosseel, Y. (2010). Detecting concealed information with reaction times: Validity and comparison with the polygraph. *Applied Cognitive Psychology*, *24*(7), 991–1002. https://doi.org/10.1002/acp.1601

Verschuere, B., & De Houwer, J. (2011). Detecting concealed information in less than a second: response latency-based measures. *Memory Detection*, *December 2015*, 46–62. https://doi.org/10.1017/cbo9780511975196.004

Verschuere, B., Kleinberg, B., & Theocharidou, K. (2015). RT-based memory detection: Item saliency effects in the single-probe and the multiple-probe protocol. *Journal of Applied Research in Memory and Cognition*, *4*(1), 59–65. https://doi.org/10.1016/j.jarmac.2015.01.001

Verschuere, B., & Meijer, E. H. (2014). What's on Your Mind? *European Psychologist*, *19*(3), 162–171. https://doi.org/10.1027/1016-9040/a000194

Visu-Petra, G., Miclea, M., & Visu-Petra, L. (2012). Reaction Time-based Detection of Concealed Information in Relation to Individual Differences in Executive Functioning. *Applied Cognitive Psychology*, *26*(3), 342–351. https://doi.org/10.1002/acp.1827

Wentura, D., & Degner, J. (2010). A practical guide to sequential priming and related tasks. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 95–116). The Guilford Press.

Yonelinas, A. P. (2002). The Nature of Recollection and Familiarity: A Review of 30 Years of Research. *Journal of Memory and Language*, *46*(3), 441–517. https://doi.org/10.1006/jmla.2002.2864

Supplementary Materials

Peer Review History

Download: https://collabra.scholasticahq.com/article/77819-guilty-on-the-go-uncovering-concealed-information-by-assessing-response-preparation-processes-in-a-go-nogo-paradigm/attachment/163193.docx?auth_token=HRv3rHaE5kLuyA-oT7I8