Importance-related Fillers Improve the Classification Accuracy of the Response Time Concealed Information Test in a Crime Scenario

Abstract

Abstract

Purpose. The Response Time Concealed Information Test (RT-CIT) can reveal when a person recognizes a relevant item among other irrelevant items, based on comparatively slower responding. Therefore, if a person is concealing knowledge about the relevance of this item (e.g., recognizing it as a murder weapon), this deception can be revealed. A recent study introduced additional “familiarity-related fillers”, and these items substantially enhanced diagnostic efficiency in detecting autobiographical data. However, the generalizability of the efficiency of fillers to other scenarios remains an open question. We empirically investigated whether new importance-related fillers enhanced diagnostic efficiency in an imaginary crime scenario.

Methods. Two hundred and thirty-nine volunteers participated in an independent samples experiment. Participants were asked to imagine either committing a crime (“guilty” group) or to imagine visiting a museum (“innocent” group). Then, all participants underwent RT-CIT testing using either a standard single probe or an enhanced single probe (with importance-related fillers) protocol.

Results. The enhanced RT-CIT (with importance-related fillers) showed high diagnostic efficiency (AUC = .810), and significantly outperformed the standard version (AUC = .562). Neither dropout rates nor exclusion criteria influenced this enhancement.

Conclusions. Importance-related fillers improve diagnostic efficiency when detecting episodic information using the RT-CIT, and seem to be useful in detecting knowledge in a wide range of scenarios.

Keywords: deception, Concealed Information Test, response time, validation

Introduction

One of the most prominent techniques used in detecting deception is the Concealed Information Test (CIT). At the end of the 1950s, Lykken (1959), proposed an autonomic nervous system-based CIT (ANS-CIT) as a scientifically-based alternative to the classic, but much criticized, technique of using polygraph interrogation with a Control Question Test (CQT; for a critical review of the usage of CQTs see, e.g., National Research Council, 2003; Ben-Shakhar, 2008). Lykken claimed that, instead of focusing on lie detection, his method focused on the detection of guilty knowledge: information regarding a certain crime detail known only to the perpetrator and the investigators[1]. To illustrate the basic notion of the CIT, let us consider a murder case scenario in which the murder weapon is known only to the offender and the Police. In this case, the CIT could include the actual murder weapon (the *probe*; e.g., "rifle") and several other weapons (*irrelevants*; e.g., "knife," and "rope"). These items would be presented sequentially to a suspect in random or sometimes pseudorandom order. Lykken and his successors showed that the CIT could be successfully used in conjunction with recording and analyzing changes in autonomic nervous system measures (see, e.g., Lykken, 1960; Verschuere, Meijer, & De Clercq, 2011; Meijer, Selle, Elber, & Ben-Shakhar, 2014). However, in the years that have elapsed since Lykken's initial publication, other measures have been introduced and used to detect the concealment of

[1]However, some studies have shown that awareness of deception on the part of the suspect can enhance test accuracy in the case of ANS-based and P300-based CITs (Ben-Shakhar & Elaad, 2003; Verschuere, Rosenfeld, Winograd, Labkovsky, & Wiersema, 2009).

information, e.g., P300 amplitude (P300-CIT; for a review, see Rosenfeld, 2020), and response time (RT-CIT; e.g., Seymour, Seifert, Shafto, & Mosmann, 2000; Verschuere, Crombez, Degrootte, & Rosseel, 2010).

In the RT-CIT, each test item is responded to with a keypress. Recognition of the probe (in the present case, "rifle") by a guilty person (who is aware of the relevance of the item) will typically result in a slower response to the item compared to irrelevant items. Therefore, based on probe-irrelevant RT differences (a RT-CIT-effect), a guilty person can be distinguished from an innocent one (Verschuere, Crombez, Degrootte, & Rosseel, 2010). The standard RT-CIT (henceforth referred to as the S-CIT) includes a single (randomly chosen) *target* irrelevant item that requires the pressing of a response key which is different from the one pressed in response to probe and nontarget irrelevant items. For example, the key "I" has to be pressed whenever the target item appears, while the key "E" has to be pressed whenever any of the other items (probe and nontarget irrelevants) appear. However, some studies have shown that the accuracy of the technique is considerably lowered or even ineffective when only one probe is used (Verschuere, Kleinberg, & Theocharidou, 2015; Kleinberg & Verschuere, 2015).

A recent study introduced a significantly enhanced version of the RT-CIT (henceforth referred to as the E-CIT) by adding filler items to the task (Lukács, Kleinberg, et al., 2017). In a knowledgeable condition of this study, the probes were certain personal details of a participant (birthdate, favorite animal, etc.)[2], which were therefore "familiar" (self-related, recognizable, etc.) to a given participant, as opposed to the irrelevants (e.g., other dates, random animal names) that were in this context rather "unfamiliar" (other-related, etc.). Two corresponding types of filler words were added to the task: (a) familiarity- and self-related filler words ("FAMILIAR," "RECOGNIZED," and "MINE") that had to be categorized using

[2] In a naive condition, the probes for participants were just other irrelevant items (random dates, animals, etc.).

the key which was also used for the target (and, thus, with the opposite key to that used for the probe and irrelevants), and; (b) unfamiliarity- and other-related filler words ("UNFAMILIAR," "UNKNOWN," "OTHER," "THEIRS," "THEM," and "FOREIGN") that had to be categorized using the key which was also used for the probe and irrelevants.

In the E-CIT it is assumed[3] that responses to familiar probes (true personal details for a given participant) should be even slower then for probes in the S-CIT because they have to be categorized *together* with unfamiliarity-related stimuli (and oppositely to familiarity-related stimuli; see, e.g., Greenwald, et al., 2009). In contrast, participants given random details as probes (i.e., not their own personal details) should see no substantial difference between probes and irrelevants (the probes should be no more familiar, self-related, or recognizable than the irrelevants), and therefore the fillers should not slow down responses to probes. The results of the Lukács, Kleinberg, et al. (2017) study confirmed this reasoning. Also, a follow-up study (Lukács & Ansorge, 2021) provided strong evidence that the meaning of the fillers matters: reversing key mapping (unfamiliarity-related fillers categorized together with targets), as well as using neutral (e.g., object name) or pseudoword fillers, both reduced, or even obliterated, the enhancing effect of fillers.

Since the meaning of fillers is therefore clearly crucial, the question arises as to how well the efficacy of fillers generalizes to other types of probe (in particular, to crime-related items), which are more likely to occur in real police investigations than autobiographical items. In the present study, we introduced importance-related fillers to the RT-CIT, to be used with crime-related items. More precisely, the fillers to be categorized together with the target were "IMPORTANT," "SIGNIFICANT," and "RECOGNIZED" – while those to be

[3] The assumption described here is only one of two key assumptions. The other assumption is less relevant to the present paper, but, in brief, it is related to the fact that the increased complexity also requires more attention throughout the task, which likely facilitates deeper processing of the stimuli (Lukács, Kleinberg, et al., 2017; see also Hu et al., 2013; Visu-Petra et al., 2013).

categorized together with the probe and irrelevants were the antonyms of these (e.g., “UNIMPORTANT”, etc.). It should be noted that these fillers are not specifically crime-related, but are related to the essence of the CIT: the importance, relevance, and recognition of probes (from the perspective of a “guilty” or knowledgeable examinee) – and therefore they are likely to be generalizable to any type of probe item.

The necessity of, and suggestions for, examining the generalizability of fillers have been mooted several times (e.g., already by Lukács, Gula, et al., 2017). But, to date, all studies except one have only used autobiographical items when examining the E-CIT. The one exception (Olson et al., 2020) included both autobiographical and crime-related probes, and, presumably due to relatively small sample sizes (ca. 19 per group), no substantial evidence was found for a difference or equivalence between crime-related and autobiographical probes, despite a general increase in probe-irrelevant reaction time differences for the authors’ P300-CIT protocol with fillers. Furthermore, there were no “innocent” control groups, and therefore it was not possible to assess whether there was an increase in diagnostic efficiency. Finally, the study involved EEG measurements, and therefore, in order to allow these measurements, the task design was necessarily slightly different from typical RT-CIT designs (in particular, it was necessary to have a longer 3 s time limit for responses). Thus the study’s findings may not be straightforwardly generalizable to the RT-CIT.

Taking into account the above-mentioned issues, the current study’s aims were: (1) to introduce importance-related fillers, and; (2) to test whether the E-CIT is more sensitive (in terms of displaying a larger probe-irrelevant mean RT difference and area under the curve) than the S-CIT in the case of episodic data.

Method

Participants

Five hundred and four participants started the experiment, but 265 participants dropped out at some point. Of the 239 participants finishing the experiment, 21 were excluded due to excessive error rates (ERs; the exclusion criterion was an ER above 50% for any item type). The final sample comprised 218 participants: 57 in an S-CIT Guilty group (age = 23.3 years, ±4.4 years [1 unknown]; 15 males, 41 females [1 unknown]); 51 in an S-CIT Innocent group (age = 22.8 years, ±3.8 years; 13 males, 37 females [1 unknown]); 53 in an E-CIT Guilty group (age = 24.7 years, ±7.7 years; 21 males, 29 females [3 unknown]); 55 in an E-CIT Innocent group (age = 22.9 years, ±5.1 years; 14 males, 41 females). Comprehensive information about dropouts and exclusions in each group is presented in Table 1.

Adult participants were recruited through Social Media and advertisements at universities and high schools. No reward was offered. All participants provided informed consent before the experiment.

Table 1

Comprehensive information about participants in each group

	Group			
	S-CIT – Innocent	S-CIT – Guilty	E-CIT Innocent	E-CIT Guilty
Initial *n*	108	131	123	142
Dropout	55 (50.9%)	70 (53.4%)	63 (51.2%)	77 (54.2%)
Exclusion – excessive error rate	2 (3.8%)	4 (6.6%)	5 (8.3%)	12 (18.5%)
Final sample	51	57	55	53

Note. Due to software limitations, initial numbers of participants include all participants, even those who only read experimental instructions. Consequently, dropout numbers reflect the numbers of participants who aborted their participation at any stage (e.g., while simply reading the initial instructions or toward the end of their experimental task). Exclusion denotes the number of participants who were excluded based on an excessive error rate (all participants). Percentages in parentheses indicate changes as compared to: (a) in the case of dropouts, all participants, and; (b) in the case of ER-based exclusions, participants remaining after dropouts.

Procedure

The experimental tasks (consideration of an imaginary scenario involving either committing a crime or visiting a museum, and participation in concealed information tests) were prepared with PsychoPy software and run on the pavlovia.org platform (Bridges et al., 2020). At the end of the experiment, participants were transferred to the Qualtrics platform to complete a memory check test[4] and personality questionnaires (unrelated to this study). The entire study was conducted in Polish.

After opening a link enclosed in an advertisement for the study, participants were transferred to a study web page containing detailed instructions. They were informed that, by continuing, they were confirming that they were at least 18 years old and that they consented to the terms of participation in the study. Next, they were randomly assigned to one of two experimental groups: guilty or innocent[5], half of the participants in each group performing the S-CIT and the rest performing the E-CIT.

Participants in the guilty groups were requested to imagine that they were committing a theft. The imaginary crime scenario was based on a study by Suchotzki et al. (2019). Participants were asked to imagine that they went to a doctor's surgery and were alone in a *waiting room*. Suddenly, they noticed that someone had left a *handbag*. They took the opportunity and looked inside it. They found a black *telephone* and decided to take it.

---

[4] Participants were asked to recall probe items. Out of all participants who completed the test, 14 did not recall all the items correctly, and 5 of these also met the criterion for exclusion due to an excessive error rate. The exclusion of the remaining 9 participants made no notable (i.e., only very small, nominal) differences to the results obtained.

[5] Of course, both guilt and innocence were only simulated, and the CIT itself implies recognition rather than guilt. Nonetheless, we consider the terms "guilty" and "innocent" to be simpler and more intuitive than alternatives such as "knowledgeable" and "naive."

Subsequently, they left the waiting room quickly. Each step of the scenario was supported with a photograph of a waiting room, a handbag, and a telephone, respectively. Subsequently, participants were provided with a summary of their actions ("You were in a *waiting room.* You took a black *telephone* from a *handbag*.") They were requested to memorize the course of the crime, especially the three crucial details (written in italics in the preceding text), which were written in capital letters in the summary. They were encouraged to repeat the details several times, and even to write them down. They were also informed that they would be tested as to whether they remembered these details correctly at the end of the study. The three crime details were used as probes in the CITs.

At the same stage of the experiment, participants in the innocent groups were requested to imagine visiting a museum. Thus, they were asked to imagine that they went to a *museum*. They were there with a couple of visitors, and suddenly they spotted a *painting* on the wall. They took the opportunity to look closer and noticed a small *village* next to a beach. Then they left the museum quickly. Again, participants in these groups were provided with a summary of their visit ("You were at a *museum*. You saw a *painting* with a small *village* next to a beach."). Further instructions were the same as for the guilty groups. However, the three details mentioned to the innocent groups were used only to ensure that all groups had similar tasks, and they were not used in any of the CITs.

Before RT-CITs, participants were informed that they were suspected of having committed the offense of theft. They were told that their task was to prove that they did not possess any information about the crime committed, and, even if they acquired some knowledge, they should hide this during their next task. Next, they were informed that they would undergo a lie detection procedure which would consist of three parts, and that each part would cover one of three details about the imaginary crime: the stolen item, the premises where the crime was committed, and the place/object from which the item was stolen (see Table 2).

After reading instructions, all groups took part in three RT-CIT blocks – one for each probe (waiting room, handbag, and telephone). The RT-CIT blocks' order was randomized within each participant. However, as previously mentioned, half of the participants from each of the guilty and innocent groups performed the S-CIT and the rest performed the E-CIT.

RT – CITs – reaction time-based Concealed Information Tests

The S-CIT. For the S-CIT, participants were informed that they would see words referring to different items in the center of their computer screen, and that their task was to categorize these as recognized or not recognized. They were told that pressing the "K" key on their keyboard meant "YES, I recognize this word" and that pressing the "D" key meant "NO, I do not recognize this word". During each part of the S-CIT, one word was chosen as a TARGET, and participants were instructed that their task was to press the "K " key ("YES, I recognize this word") each time they saw it on the screen. In the case of the rest of the words, participants were told to press the "D" key ("NO, I do not recognize this word").

Before each part (every S-CIT), participants completed three practice rounds. During the first two practice rounds, participants saw the question: "Do you recognize this word?" on the top of their screen, and response guidance: "D – NO" on the left part of the screen, and "K – YES" on the right. This information was not presented during the third practice round and the main task. During the first practice round, each word was presented without a time limit: the next word was not presented until participants pressed either key. If participants gave an incorrect response, the message "WRONG" written in red was shown for 200 ms, after which the next stimulus was presented. During the second round, the time limit for responding was 1,500 ms. If participants did not respond within this time, "TOO SLOW" written in red was presented below the stimulus for 200 ms. The inter-stimulus interval was randomized within the range 500 to 1,000 ms (in all practice rounds and in the main task). If participants failed to complete either of the first or second practice rounds correctly, they were not repeated. During the third practice round, the response time limit was shortened to 1,000 ms, as it was in the subsequent main task, and this round was repeated twice (regardless of a participant's error rate). In each practice round, all six of the stimuli (probe, target, and 4 irrelevant items – see Table 2) were randomly presented once. During the main task, each stimulus was presented 20 times – 120 stimuli for each part (20 x probe, 20 x target, 80 x irrelevant – 4 irrelevant items 20 times each). The order of stimuli was randomized in groups of six (one probe, one target, and four irrelevant).

The E-CIT. Instructions for the enhanced version were very similar to the S-CIT. However, instead of categorizing words as recognized or not, participants were informed that they should categorize them as important (by pressing the "K" key) or unimportant (by pressing the "D" key). They were told that they should press "K " for the target word as well as important fillers, and press "D" for other words including unimportant fillers. Other information was identical to that for the S-CIT. As to the practice rounds and main tasks, the only differences concerned: (1) questions presented during the first two practice rounds ("Is it significant for you?" instead of "Do you recognize this word?"), and; (2) the number of items presented during practice rounds and main tasks. Here, in each practice round participants saw each stimulus once in a random order (1 probe, 1 target, 4 irrelevant items, 3 important fillers, 6 unimportant fillers – 15 in total). During each main task, the probe, target and each irrelevant item were presented 20 times (120 presentations in total), and, additionally, each important and unimportant filler was presented six times (54 presentations in total; see also Tables 2 and 3). The order of stimuli was pseudorandom – within each stimulus category, the order was prepared following the restrictions proposed by Lukács, Kleinberg, et al. (2017), and these same fixed trial orders for each category were used for all participants (but with categories in random order).

Table 2

Items used in the RT-CIT

	Category		
	Stolen item	Premises	Container
Probe	Telephone (*Telefon*)	Waiting room (*Poczekalnia*)	Handbag (*Torebka*)
Irrelevants	Wallet, Money, Camera, Documents (*Portfel, Pieniądze, Aparat, Dokumenty*)	Bathroom, Shop, Cafe, Cloakroom (*Łazienka, Sklep, Kawiarnia, Szatnia*)	Basket, Briefcase, Jacket, Suitcase (*Koszyk, Teczka, Kurtka, Walizka)*
Target	Computer (*Komputer*)	Office (*Biuro*)	Backpack (*Plecak*)
Important fillers (only E-CIT)	Important, Significant, Known (*Ważny, Istotny, Znany*)		
Unimportant fillers (only E-CIT)	Unimportant, Insignificant, Unknown, Regular, Neutral, Unfamiliar (*Nieważny, Nieistotny, Nieznany, Zwykły, Obojętny, Obcy*)		

Note: The Polish words that were actually used in the experiment are shown in parentheses.

Table 3

The number of stimuli presented per item type for each RT-CIT version.

	S-CIT	E-CIT
Probes	20	20
Irrelevant	80	80

Target	20	20
“Important” filler	-	18
"Unimportant” filler	-	36
TOTAL	120	192

Results

The main measure compared during analyses was the RT-CIT effect: the difference in reaction times to probe and irrelevant stimuli. Only data for errorless trials were analyzed, and trials with reaction times below 150 ms or above 1,000 ms were also excluded. Bayes factors (BFs) are reported using the default r-scale of 0.707, and, in the case of ANOVAs, inclusion BFs are based on matched models. All analyses were performed in R (R Core Team, 2019) using packages designed by Kelley (2018), Morey and Rouder (2018), and Robin et al. (2011). All tests were preregistered except where explicitly noted otherwise.

Mean reaction times, mean error rates and classification accuracies are presented in Table 4.

Table 4

Response Times and AUCs

	Probe	Irrelevant	Target	Filler-NT	Filler-T	P – I	AUC [95% CI]	TPR	TNR
S-CIT RT									
Guilty	405±54	404±49	485±48			1.1±16.1	.562	.42	.78
Innocent	404±52	407±47	494±71			-2.8±14.0	[.453, .671]		
E-CIT RT									
Guilty	522±73	506±63	579±59	543±78	621±55	16.0±26.9	.810	.93	.62
Innocent	490±59	496±55	571±50	545±71	618±52	–6.3±15.8	[.728, .891]		

S-CIT ER									
Guilty	1.7±3.3	1.6±3.5	19.7±14.5			0.08±1.85	.529	.25	.88
Innocent	1.1±1.7	1.2±2.0	20.3±12.3			–0.05±1.28	[.420, .638]		
E-CIT ER									
Guilty	2.1±3.2	3.8±4.8	22.9±18.8	5.8±5.0	30.3±23.4	–1.70±3.49	.231		
Innocent	2.7±5.1	2.2±5.2	21.9±15.7	7.1±5.3	28.0±18.0	0.49±1.43	[.143, .319]		

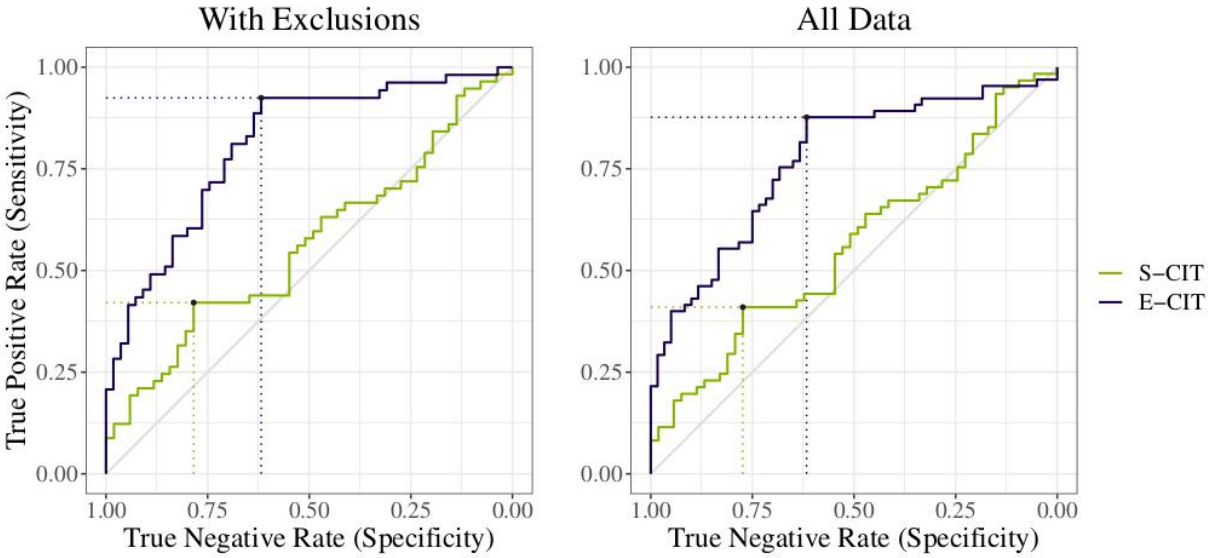
Note. Means and *SD*s for individual RT means (ms) for different item types, and for probe-irrelevant differences (P – I), and, most importantly, corresponding AUCs (95% CIs in brackets). TPR: true positive rates (ratio of correctly detected 'guilty' participants), TNR: true negative rates (ratio of correctly detected 'innocent' participants), using arbitrary optimal cutoffs (maximal Youden's index) for classification. Filler-T: "target-side" important fillers; Filler-NT: "nontarget-side" unimportant fillers. AUC: area under the curve.

Reaction times

A two-way between-subjects ANOVA (condition: guilty vs. innocent, and RT-CIT protocol: S-CIT vs E-CIT) on probe-irrelevant RT mean differences showed significant main effects of condition, $F(1, 212) = 33.93$, $p < .001$, $\eta_p^2 = .138$, 90% CI [.073, .209], $\eta_G^2 = .138$, $BF_{10} = 1.54 \times 10^5$, and RT-CIT protocol, $F(1, 212) = 7.92$, $p = .005$, $\eta_p^2 = .036$, 90% CI [.006, .085], $\eta_G^2 = .036$, $BF_{10} = 4.66$. Most importantly, there was a significant interaction between Condition and RT-CIT protocol, $F(1, 212) = 18.04$, $p < .001$, $\eta_p^2 = .078$, 90% CI [.030, .141], $\eta_G^2 = .078$, $BF_{10} = 562.18$. A follow-up one-sided *t*-test (as preregistered) comparing protocols in guilty groups showed that, as expected, probe-irrelevant differences in the E-CIT were significantly larger than in the S-CIT, with a difference of 16.8 ms, 90% CI [11.7, ∞], $t(91.3) = 4.28$, $p < .001$, $d = 0.83$, 90% CI [0.50, ∞], $BF_{10} = 1189.29$. Correspondingly, areas under the curve (AUC) for the E-CIT proved significantly larger than those for the S-CIT via a one-sided DeLong test, with a difference of .247, 95% CI [.109, .385], $D(198.4) = 3.56$, $p < .001$ (see Figure 1).

Figure 1

Receiver operating characteristic curves



Note. True positive and true negative rates at all possible cutoff points for S-CIT and E-CIT probe-irrelevant RT mean predictors, for data after ER-based exclusions (left panel), and for all data without any exclusions (right panel).

Error rates

For probe-irrelevant ER differences, an analogous two-way between-subjects ANOVA showed a significant main effect of condition, $F(1, 212) = 16.37$, $p < .001$, $\eta_p^2 = .072$, 90% CI [.025, .133], $\eta_G^2 = .072$, $BF_{10} = 133.59$, but not of RT-CIT protocol, $F(1, 212) = 3.68$, $p = .056$, $\eta_p^2 = .017$, 90% CI [0, .056], $\eta_G^2 = .017$, $BF_{01} = 1.37$. While there was a significant interaction between Condition and RT-CIT protocol, $F(1, 212) = 17.05$, $p < .001$, $\eta_p^2 = .074$, 90% CI [.027, .136], $\eta_G^2 = .074$, $BF_{10} = 438.27$, surprisingly, this interaction was the reverse of what was expected. Thus, while the difference between probe and irrelevant ERs was practically negligible for the S-CIT, in the case of the E-CIT the ERs for irrelevants were higher than for probes (and were therefore also not diagnostic of guilt in the conventional probe-minus-irrelevant form; see Table 4). However, ERs were generally very low (all below

3%), and therefore a substantial speed-accuracy tradeoff seems unlikely. In any case, ERs are not generally used as diagnostic data in the RT-CIT.

Exploratory analyses: dropout rate differences and adjusted exclusion criteria

For the present data, Pearson's chi-square tests and independent multinomial contingency table BFs both provided evidence for the similarity of dropout rates between the E-CIT and S-CIT protocols when only the guilty conditions were included, with a nominal difference of a mere 0.8%, 95% CI [–13.3, 11.8], χ^2 (1) < .01, p = .993, BF_{01} = 6.60, and when both the guilty and innocent conditions were included, the difference here being 0.5%, 95% CI [–9.7, 8.6], χ^2 (1) < .01, p = .977, BF_{01} = 8.93.

All ER-based exclusions in the present study were due to ERs that were too high for targets or importance-related fillers. But, even for these, none of the participants had ERs that were high enough to indicate a complete absence of attention (e.g., zero correct responses). Therefore, we repeated all the analyses with all 239 participants who completed the tests included (i.e., with no ER-based exclusions). All the results remained statistically significant with no notable changes. In the interests of brevity, here we report only the results for the most relevant test: the AUC comparison. Thus, for the S-CIT without exclusions the AUC was .561, 95% CI [.455, .667], while for the E-CIT without exclusions the AUC was .773, 95% CI [.690, .855], with DeLong's test again showing very strong evidence for a difference (.211, 95% CI [.075, .347]), D(218.8) = 3.08, p = .001 (see Figure 1).

Discussion

Deception detection researchers will probably always seek methods which are effective, useful and easy to conduct in practice. It is hard to imagine a more useful or easier method than the RT-CIT since it requires only a computer or just a smartphone (Lukács, Kleinberg, Kunzi, & Ansorge, 2020). However, as mentioned in the Introduction, the effectiveness of the RT-CIT is still debatable. But the current study allowed us to establish at least three important findings: (1) adding importance-related fillers significantly improves RT-CIT accuracy in the case of episodic information; (2) in contrast to the standard version (S-CIT), the enhanced RT-CIT (E-CIT) can be effective even where information is of low salience, and; (3) there are no substantial differences in dropout between the two versions of the RT-CIT, and the imposition of less strict exclusion criteria does not affect the main results obtained.

Importance-related fillers significantly improve RT-CIT accuracy

In this study we again demonstrated the enhancing influence of filler items, and we successfully introduced importance-related fillers, which seem applicable in a broad range of situations, such as testing for both autobiographical and episodic information. The E-CIT with importance-related fillers clearly outperformed the S-CIT where the CIT-effect and accuracy were concerned. Thus, importance-related fillers seem to have strong advantages over familiarity-related fillers from a practical point of view. Relative to familiarity-related fillers, importance-related fillers (important, significant, and recognize) may be less misleading, more intuitive, and easier to explain (for examiners) and follow (for participants) in crime related cases than familiarity-related fillers. From a semantic point of view, in criminal cases concerning theft the probe item is not MINE (which is a familiar filler), but probe could belong to THEM or be THEIR (which are unfamiliar fillers). At the same time, all kinds of probes are IMPORTANT, SIGNIFICANT or RECOGNIZED, and none of them are UNIMPORTANT, NEUTRAL etc.

It is necessary to conduct future studies comparing importance-related and familiarity-related fillers. It is likely that, while an E-CIT with familiarity-related fillers may be just as accurate as an E-CIT with importance-related fillers in any scenario, despite the moderate differences in meaning, importance-related fillers will have a clear advantage because of their greater usefulness and their potentially wider applicability.

The enhanced RT-CIT can be effective even for information of low salience.

The classification accuracy for S-CIT was not much greater than would have been expected by chance. This is somewhat surprising, although not unprecedented in case of the S-CIT. Some previous studies have shown that the effectiveness of a standard single probe RT-CIT may be unsatisfactory, especially when attempting to detect the possession of information which is of low salience (Verschuere, Kleinberg, & Theocharidou, 2015; Kleinberg & Verschuere, 2015).

The imaginary crime scenario used in our research was based on a study by Suchotzki et al. (2019). In this scenario, no crime is actually committed, or role played. Participants only imagine performing an action described by their instructions, and are asked to memorize certain details. From a participant's point of view this information is only relevant because it is part of an experimental task. It is not personally meaningful (unlike their date of birth or name), and it can be assumed that it is not associated with high, or even moderate, emotional arousal, unlike information about a real crime that has been committed. Even compared to information acquired during a mock crime, where participants perform real actions (for example, touching and feeling objects), information about an imaginary crime can be considered to have very low salience. Furthermore, our experiment was conducted online, and participants' engagement may have been quite low. To some extent, these possibilities are confirmed by the fact that 14 of the participants who completed the whole experimental procedure failed to recall the probes correctly. It is highly probable that the standard single probe RT-CIT was rather ineffective in detecting the type of knowledge involved because of the probes' low salience.

Higher exclusion rates are not important for enhanced results

Our third important finding concerns dropout and exclusion rates. Some authors have stressed that the E-CIT has a notable limitation in that higher dropout and ER-based exclusion rates are observed compared to the S-CIT and other protocols (Koller et al., 2020; Suchotzki et al., 2018). However, these observations have been made exclusively from online studies. In a laboratory study, Olson et al. (2020) did not observe a single dropout or ER-based exclusion among a total of 38 participants in their E-CIT group, and they still successfully replicated the finding of larger probe-irrelevant RT mean differences in this group compared to their S-CIT group. Nonetheless, since their task design was somewhat different because they took concurrent EEG measurements (in particular there was a longer response time limit than is typical for the RT-CIT: 3 s instead of 0.8-1.5 s), it is not certain that this finding applies to the RT-CIT in general.

The results of this online study showed that, unlike the previous online studies (e.g., Lukács, Kleinberg, et al., 2017), the dropout rate did not differ significantly between the two protocols. It is very probable that the differences in dropout rate in the former studies was due to their very strict criteria for passing practice round in E-CIT, especially in comparison to those for their S-CIT group.

Furthermore, in the previous studies (and also in this one) exclusion criteria for excessive error rates were all arbitrarily set at a "minimum 50% ER for each item type," which is far stricter for the E-CIT (5 different items types) than for the S-CIT (only 3 different item types). Nevertheless, in the current study, we were able to show that even without exclusion of participants with excessive error rates, the results remained almost the same. To provide even more proof in this regard, we reanalyzed the S-CIT and E-CIT data of Lukács, Kleinberg, et al. (2017; Experiment 1 only – since the other experiments contained no "innocent" control data), and no matter which specific exclusion criterion we chose, the key results were the same. To illustrate this, for simplicity we again report the outcomes when all participants completing the tests are included in an analysis ($N = 314$). For the 2017 S-CIT data without exclusions the AUC was .627, 95% CI [.548, .706], while for the 2017 E-CIT data without exclusions the AUC was .849, 95% CI [.780, .918], DeLong's test once again showing very strong evidence for a difference, $D(310.2) = 4.14$, $p < .001$. [6]

Limitations and further studies

At least two limitations of the present study should be noted. First, the study employed an imaginary crime scenario. While this was advantageous in that it allowed a study involving episodic information to be conducted via the Internet, it has to be acknowledged that an imaginary crime scenario is probably less similar to a real crime than mock-crime scenarios. Second, as mentioned in the Method section, in the case of the E-CIT (in contrast to the S-CIT) we presented the stimuli in each block in a pseudorandom order. While it is unlikely that this had any substantial influence on the results obtained, it should be borne in mind as a potential limitation. Furthermore, we used only new, importance-related fillers, so it was impossible to verify whether these were less, more, or equally efficient than familiarity-related fillers in the scenario presented. This would be an interesting issue for further studies to explore.

Conclusions

[6] A more thorough discussion on why dropout and exclusion rates make no difference regarding the enhancement can be found in the Online Appendix via https://osf.io/5thpe/.

The practical implication of increasing the diagnostic efficiency of the RT-CIT is straightforward: it would lead to more correct classification of real-world suspects as to whether they have recognized a given probe item. While familiarity-related fillers have been repeatedly shown to provide such an enhancement in efficiency for autobiographical details, the present paper is the first to show that importance-related fillers can also be of great benefit, and shows that such fillers are highly generalizable to different types of scenario. Which types of fillers work best (e.g., more specific, crime-related fillers, such as "stolen," or more general fillers, such as those used in the present study) remains to be explored in future research.

We have also shown, for the first time, that neither dropout rates nor ER-based exclusions pose a limitation to the enhancement in efficiency achieved. Thus, to our knowledge, the E-CIT, while repeatedly proven to outperform the S-CIT, has no notable limitations in comparison to any other RT-CIT protocol.

References

Ben-Shakhar, G., & Elaad, E. (2003). The validity of psychophysiological detection of information with the Guilty Knowledge Test: A meta-analytic review. *Journal of Applied Psychology*, *88*(1), 131.

Ben-Shakhar, G. (2008). The case against the use of polygraph examinations to monitor post-conviction sex offenders. *Legal and Criminological Psychology*, *13*(2), 191-207.

Bridges, D., Pitiot, A., MacAskill, M. R., & Peirce, J. W. (2020). The timing mega-study: Comparing a range of experiment generators, both lab-based and online. *PeerJ, 8*, e9414. https://doi.org/10.7717/peerj.9414

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology, 97*(1), 17–41. https://doi.org/10.1037/a0015575

Hu, X., Evans, A., Wu, H., Lee, K., & Fu, G. (2013). An interfering dot-probe task facilitates the detection of mock crime memory in a reaction time (RT)-based concealed information test. *Acta Psychologica, 142*(2), 278–285. https://doi.org/10.1016/j.actpsy.2012.12.006

Kelley, K. (2018). *MBESS: The MBESS R Package.* R package version 4.4.3. https://CRAN.R-project.org/package=MBESS

Kleinberg, B., & Verschuere, B. (2015). Memory detection 2.0: The first web-based memory detection test. *PloS one*, *10*(4), e0118715.

Koller, D., Hofer, F., & Verschuere, B. (2020). *Different target modalities improve the single-probe protocol of the response time-based Concealed Information Test*. https://doi.org/10.31219/osf.io/385pd

Lukács, G., & Ansorge, U. (2021). The mechanism of filler items in the response time concealed information test. *Psychological Research*. https://doi.org/10.1007/s00426-020-01432-y

Lukács, G., Gula, B., Szegedi-Hallgató, E., & Csifcsák, G. (2017). Association-based Concealed Information Test: A novel reaction time-based deception detection method. *Journal of Applied Research in Memory and Cognition*, *6*(3), 283–294. https://doi.org/10.1016/j.jarmac.2017.06.001

Lukács, G., Kleinberg, B., Kunzi, M., & Ansorge, U. (2020). Response time concealed information test on smartphones. *Collabra: Psychology*, *6*(1).

Lukács, G., Kleinberg, B., & Verschuere, B. (2017). Familiarity-related fillers improve the validity of reaction time-based memory detection. *Journal of Applied Research in Memory and Cognition*, *6*(3), 295–305. https://doi.org/10.1016/j.jarmac.2017.01.013

Lykken, D. T. (1959). The GSR in the detection of guilt. *Journal of Applied Psychology*, *43*(6), 385.

Lykken, D. T. (1960). The validity of the guilty knowledge technique: The effects of faking. *Journal of Applied Psychology*, *44*(4), 258.

Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of Bayes factors for common designs* (Version 0.9.12-4.2). Retrieved from https://CRAN.R-project.org/package=BayesFactor

National Research Council (2003). The polygraph and lie detection. Committee to review the scientific evidence on the polygraph. Division of behavioral and social sciences and education. Washington, DC: The National Academies Press

Olson, J., Rosenfeld, J. P., & Perrault, E. (2020). Familiarity-related filler items enhance the RT CIT, but not the P300 CIT: Episodic vs. Semantic memory protocols. *International*

Journal of Psychophysiology, 158, 370–379. https://doi.org/10.1016/j.ijpsycho.2020.10.001

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*(1), 77. https://doi.org/10.1186/1471-2105-12-77

Rosenfeld, J. P. (2020). P300 in detecting concealed information and deception: A review. *Psychophysiology*, *57*(7), e13362.

Seymour, T. L., Seifert, C. M., Shafto, M. G., & Mosmann, A. L. (2000). Using response time 668 measures to assess "guilty knowledge". Journal of Applied Psychology, 85(1), 30–37. 669 https://doi.org/10.1037//0021-9010.85.1.30

Suchotzki, K., De Houwer, J., Kleinberg, B., & Verschuere, B. (2018). Using more different and more familiar targets improves the detection of concealed information. *Acta Psychologica, 185*, 65–71. https://doi.org/10.1016/j.actpsy.2018.01.010

Suchotzki, K., Kakavand, A., & Gamer, M. (2019). Validity of the reaction time Concealed Information Test in a prison sample. *Frontiers in psychiatry*, *9*, 745.

Verschuere, B., Crombez, G., Degrootte, T., & Rosseel, Y. (2010). Detecting concealed information with reaction times: Validity and comparison with the polygraph. *Applied Cognitive Psychology*, *24*(7), 991-1002.

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.

Verschuere, B., Meijer, E., & De Clercq, A. (2011). Concealed information under stress: A test of the orienting theory in real-life police interrogations. *Legal and criminological psychology*, *16*(2), 348-356.

Verschuere, B., Rosenfeld, J. P., Winograd, M. R., Labkovsky, E., & Wiersema, R. (2009). The role of deception in P300 memory detection. *Legal and Criminological Psychology*, *14*(2), 253-262.

Visu-Petra, G., Varga, M., Miclea, M., & Visu-Petra, L. (2013). When interference helps: Increasing executive load to facilitate deception detection in the Concealed Information Test. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00146