Received: 28 February 2020 | Revised: 30 July 2020 | Accepted: 20 October 2020

DOI: 10.1002/acp.3759

WILEY

RESEARCH ARTICLE

Addressing selective attrition in the enhanced response time-based concealed information test: A within-subject replication

Gáspár Lukács

University of Vienna, Vienna, Austria

Correspondence
Gáspár Lukács, Faculty of Psychology, University of Vienna, Liebiggasse 5, A-1010 Vienna, Austria.
Email: gaspar.lukacs@univie.ac.at

Summary

The response time-based concealed information test can reveal when a person recognizes a relevant item among other, irrelevant items, based on comparatively slower responding. Thereby, if a person is concealing the knowledge about the relevance of this item (e.g., recognizing it as a murder weapon), this deception can be revealed. A recent study, conducted online and using a between-subject design, introduced a significantly enhanced version by including additional items in the task. While this modified version outperformed the original version, it also resulted in a much higher rate of participant dropouts (i.e., participants leaving the experiment's website without completing the task). The grave implication is that the perceived enhancement is perhaps merely due to selective attrition. Therefore, the current experiment replicates the original one, but using a within-subject design. The results show that there is a large enhancement even when selective attrition is prevented.

KEYWORDS

concealed information test, deception, replication, response time, selective attrition, validation

1 | INTRODUCTION

The response time-based concealed information test (RT-CIT) aims to reveal whether a person is concealing knowledge regarding a certain detail. To illustrate the CIT, let us consider a murder case scenario in which the murder weapon is known only to the perpetrator and the investigators. 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 sequentially presented to a suspect in random order. When each item has to be responded to with a keypress, the recognition of the probe (in this case, "rifle") by a guilty person (who is aware of the relevance of that item) will typically result in a slower response to that item than to the irrelevant items. Thereby, based on the probe-irrelevant RT differences, a guilty person can be distinguished from innocent ones.

The standard CIT (S-CIT) includes a single (randomly chosen) target irrelevant item that requires pressing a response key different from the response key for 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, a recent study introduced a significantly enhanced version (E-CIT) by adding filler items to the task (Lukács, Kleinberg, & Verschuere, 2017). In that study, the probes were the participants' certain personal details (birthday, favorite animal, etc.), which were therefore "familiar" (self-related, recognizable, etc.) to the given participant, as opposed to the irrelevants (e.g., other dates, random animal names) that were in this respect rather "unfamiliar" (other-related, etc.). Two corresponding kinds of fillers were added to the task: (a) familiarity-referring words ("FAMILIAR," "RECOGNIZED," and

"MINE") that had to be categorized with the same key as the target (and, thus, with the opposite key than the probe and the irrelevants), and (b) unfamiliarity-referring words ("UNFAMILIAR," "UNKNOWN," "OTHER," "THEIRS," "THEM," and "FOREIGN") that had to be categorized with the same key as the probe and irrelevants. It was assumed that responses to the familiar probes (the true personal details of the given participant) would be even slower because they have to be categorized together with unfamiliarity-referring expressions (and opposite to familiarity-referring expressions; see e.g., Greenwald, Poehlman, Uhlmann, & Banaji, 2009). In contrast, participants with random details as probes (i.e., not their own personal details) see no substantial difference between probes and irrelevants (in particular: the probes are no more familiar, self-related, or recognizable, than the irrelevants), and therefore the fillers do not slow down the responses to the probe.

The other important assumption was that the increased cognitive load due to the increased complexity of the test (more items, etc.) required more attention throughout the task, which likely facilitated deeper semantic processing of the stimuli (Lukács et al., 2017, p. 3; see also Visu-Petra, Varga, Miclea, & Visu-Petra, 2013). More specifically, in a CIT without filler items (and with a single probe and single target; Verschuere, Kleinberg, & Theocharidou, 2015), examinees have to look out for a single target item, pressing the alternative key for all other items. Thereby, they may focus on the target and to some extent ignore the content and meaning of the rest of the items. By adding fillers (to be categorized with the same key as the target), examinees have to pay more attention, and are induced to more carefully process the content and meaning of each of the presented items. This, in turn, also increases the attention to the relevant probe (which otherwise might be ignored to some extent), resulting in even slower keypress responses to it (in case of a recognized probe, such as by a guilty examinee).

The study did not provide any detailed investigation of the underlying fundamental processes—nonetheless, the enhancing effect of fillers was plainly demonstrated: the rate of correctly detected "guilty" and "innocent" (as simulated in the study) participants rose from .68 to .94, as measured by areas under the curves (e.g. National Research Council, 2003, pp. 342–344).

1.1 | Selective attrition

Pertinent to the present study, there was an apparent limitation in the study of Lukács et al. (2017). Due to the increased complexity of the task, a substantial ratio of the online participants dropped out (left the experiment's website without completing the task) when using the E-CIT (10.0–11.4%; though also 3.5–8.5% in S-CIT), as estimated by comparing the number of first practice round completions with full test completions. The factual, precise dropout rates were reported by Lukács and Ansorge (2019), who also recorded all participant who began the practice task at all (starting at the instruction page). The difference in this case was striking: 60–63% dropout in E-CIT, while only 18–19% in S-CIT.

The very grave implication is that the enhancement in the E-CIT is perhaps merely an artificial construct due to selective attrition (Zhou & Fishbach, 2016): Participants who dropped out when using the E-CIT (but not when using the S-CIT) may have been the ones that are generally less susceptible to the CIT (i.e., would have smaller probe-irrelevant differences). For example, one reason could be that there are participants who are generally less motivated and pay little attention to their tasks. These participants would decide to drop out when they are to perform the E-CIT because the task is too complex for them, while they would complete the simpler S-CIT task, but with suboptimal results. Thereby, it would appear that the E-CIT outperforms the S-CIT, while the larger effect in case of the E-CIT would be merely due to different types of participants having performed it.

If this were true (i.e., if the findings of Lukács et al., 2017, were a mere artifact), then not only would all the important practical implications be wrong, but all follow-up studies and considerations would be based on a fallacy. Indeed, already several follow-up studies have been published (e.g., Lukács, Grządziel, Kempkes, & Ansorge, 2019; Olson, Rosenfeld, & Perrault, 2019; Suchotzki, De Houwer, Kleinberg, & Verschuere, 2018), and even more are in progress. Therefore, assessing this potential confound is an urgent matter.

To address this concern, the current study replicated the original experiment, but using a within-subject design, with each participant performing both the S-CIT and the E-CIT. The procedure was identical[1] to Experiment 1 in Lukács and Ansorge (2019; with only minor differences compared to the study by Lukács et al., 2017), except that only the S-CIT and the E-CIT conditions were measured, both with simulated guilty participants, using a within-subject design.

2 | METHOD

The methods and analyses were preregistered at https://osf.io/9q8gk (analyses that were not preregistered are under the heading "Exploratory Tests"). All collected data is available via https://osf.io/wu5cf.

2.1 | Participants

The experiment was run on Figure Eight (www.figure-eight.com), an online crowdsourcing platform where participants from anywhere in the world can register to complete small online tasks (Kleinberg & Verschuere, 2015; Peer, Samat, Brandimarte, & Acquisti, 2015). Only subjects who had taken other experiments on the website seriously ("Level 3 contributors") were invited for the present experiment.

The task was completed by 61 participants[2]: 29 starting with S-CIT ($M_{age} \pm SD_{age}$ = 34.4 ± 9.1; 79.3% male; 19 [39.6%] dropouts), and 32 starting with E-CIT ($M_{age} \pm SD_{age}$ = 33.4 ± 8.4; 56.3% male; 29 [47.5%] dropouts).

All participants performed both the S-CIT and the E-CIT, in random order. All participants simulated guilty suspects: the probes were the participants' self-reported autobiographical identity details (e.g., their country of origin).

2.2 | Procedure

Before beginning the experiment, all participants agreed to the informed consent to proceed further. Participants then provided demographic information, and selected their three autobiographical details: country of origin, date of birth (month and day), and favorite animal. This was followed by the very short (3 min) LexTALE English competency test (Lemhöfer & Broersma, 2012), in which 60 words are presented, among which 40 are real English words, while 20 are nonwords, and the instruction is to decide, for each word, whether it is an actual English word. This test was implemented as described at www.lextale.com, with the only difference that a 4-s time limit applied to each response to curb possible cheating (i.e., looking up the words online or in a dictionary during the task). The LexTALE minimum score for upper intermediate (B2) level is 60% accuracy (Lemhöfer & Broersma, 2012, p. 341). Consequently, those who did not achieve a score above a more lenient threshold of 55% clearly did not have the required English skill, and therefore were automatically disqualified and redirected to the Figure Eight website. Then followed the first of the CITs as described below.

2.2.1 | Item selection

Participants were informed that the following task simulates a lie detection scenario, during which they should try to hide their identities. They were then presented a short list of randomly chosen items within each of the three categories in the task (countries, dates, animals).[3] The participants were asked to choose any (but a maximum of two per category) items that were personally meaningful to them or in any way appeared different from the rest of the items on those lists. Subsequently, the four irrelevants and one target were randomly selected from among the non-chosen items, in each category.

Next, participants were shown their three targets, and were asked to memorize these items to recognize them as requiring a different response during the following task. On the next page, participants were asked to recall the memorized items, and could proceed only if they selected these items correctly from a dropdown menu. If any of the selected items was incorrect, the participant received a warning and was redirected to the previous page.

2.2.2 | S-CIT task

During the S-CIT, items were presented one by one in the center of the screen, and participants had to categorize them by pressing one of two keys ("E" or "I") on their keyboard. Each block contained six different items, which were presented repeatedly in random order (see Test Structure): the probe, the four irrelevants, and the target. Participants had to press "I" whenever the target appeared, while they had to press "E" whenever the probe or an irrelevant appeared.

2.2.3 | E-CIT task

The E-CIT differed from the S-CIT only in that it also contained filler items. Whenever a familiarity-referring filler appeared ("FAMILIAR," "RECOGNIZED," and "MINE") participants had to press the "I" key (same as for targets), while whenever an unfamiliarity-referring filler appeared ("UNFAMILIAR," "UNKNOWN," "OTHER," "THEIRS," "THEM" and "FOREIGN"), they had to press "E" (same as for probe and irrelevants).

2.2.4 | Test structure

The inter-trial interval always randomly varied between 400 and 700 ms. In case of an incorrect response or no response within the given time limit (800 ms, except for practice, see below), the caption "WRONG" or "TOO SLOW" appeared in red color, respectively, below the stimulus for 400 ms, followed by the next trial.

The main task was preceded by a comprehension check and two practice tasks. The check served to ensure that the participant had fully understood the task. The items consisted of 21 randomly ordered trials, including the 12 different main items and each of the 9 possible fillers. During the comprehension check, participants had plenty of time (10 s) to choose a response. However, each trial required a correct response. In case of an incorrect response, the participant was reminded of the instructions and had to repeat this check.

In the following first practice task, the response window was longer than in the main task (2 s instead of 800 ms), while the second practice task had the same design as the main task. Both practice tasks consisted of 14 trials, in a way that two successive tasks always contained all of the possible items in the task (3 probes, 12 irrelevants, 3 targets, 9 fillers). In either practice task, in case of too few valid responses, the participants were reminded of the instructions, and had to repeat the practice task. The requirement was a minimum of 60% valid responses (correct key between 150 and 800 ms) for each of the following item types: targets; familiarity-referring fillers; unfamiliarity-referring fillers; main items (probes and irrelevants together).

These initial practice tasks always corresponded to the CIT version that came first (S-CIT or E-CIT). After the first CIT, there was a last practice round with items corresponding to the upcoming second CIT.

The main task, in each test, contained three blocks, one for each category (countries, dates, or animals; in random order). In each block, each probe, irrelevant, and target was repeated 18 times. Within each of the three blocks, the order of the items was randomized in groups: first, all six items (one probe, four irrelevants, and one target) in the given category were presented in a random order, then the same six items were presented in another random order (but with the restriction that the first item in the next group was never the same as the last item in the previous group).

For the E-CIT only, fillers were placed among these items in a random order, but with the restrictions that an filler trial was never

followed by another filler trial, and each of the 9 fillers (3 familiarity-referring, 6 unfamiliarity-referring) preceded each of the 3 probes, 3 targets, and 12 irrelevants exactly one time.

2.3 | Data analysis

For all analyses, RTs below 150 ms were excluded. For RT analyses, only correct responses were used. Accuracy was calculated as the number of correct responses divided by the number of all trials (after the exclusion of those with RT below 150 ms).

To demonstrate the magnitude of the observed effects for *F* tests, partial eta-squared (η_p^2) values are shown along with their 90% CIs (Steiger, 2004). For *t* tests, Welch-corrected statistics are reported for parametric tests (Delacre, Lakens, & Leys, 2017), Wilcoxon statistics for nonparametric tests, and, to demonstrate the magnitude of the observed effects, Cohen's *d* values as standardized mean differences and their 95% CIs (Lakens, 2013). All analyses were conducted in R (R Core Team, 2019; via: Kelley, 2018; Lawrence, 2016).

3 | RESULTS

Aggregated means of RT means and accuracy rates, for the different stimulus types, for S-CIT and for E-CIT, from the present within-subject study as well as from the original between-subject study (Lukács et al., 2017), are given in Table 1.

Note: Means and *SD*s (in the format of *M* ± *SD*) for individual RT means (RTs, in ms), and accuracy rates (ARs, in percentages); for Probe (participants' personal item), Irrelevant (non-personal items), Target (designated non-personal item that requires different response), Filler-F: Familiarity-referring fillers, Filler-U: Unfamiliarity-referring fillers, P–I (individual probe-minus-irrelevant values), SMD_{P-I}: standardized mean difference ("Cohen's *d*") with 95% CIs; per version (S-CIT and E-CIT), and per research design (the present within-subject design versus the original between-subject design in the study by Lukács et al., 2017).

Successfully replicating the previous results, the probe-irrelevant RT mean differences proved to be larger in the E-CIT than in the S-CIT, $t(60) = 8.35$, $p < .001$, $d_{within} = 1.07$, 95% CI [0.75, 1.38]. Analogously to RT means, the probe-irrelevant accuracy rate differences proved to be larger in the E-CIT than in the S-CIT, $W = 1197$, $p = .001$, $d_{within} = 0.44$, 95% CI [0.18, 0.70].

These results were also directly compared to the corresponding results by Lukács, Kleinberg, et al. (2017; merging all three experiments, using the raw data available at https://osf.io/kv65n), in an ANOVA with the factors version (S-CIT vs. E-CIT) and research design (the between-subject design vs. the within-subject design; i.e., the results from the original study vs. the present results); see Figure 1, left panel. The between-subject design resulted in somewhat higher probe-irrelevant differences in general than the present within-subject design, $F(1,605) = 11.83$, $p = .001$, $\eta_p^2 = .019$, 90% CI [.005, .041]. More importantly, the interaction was not significant, although there is a weak indication that the difference between the E-CIT and S-CIT may be slightly smaller in case of the within-subject design, $F(1,605) = 2.88$, $p = .090$, $\eta_p^2 = .005$, 90% CI [0, .018]. That is to say, the original between-subject design may have indeed slightly overestimated

TABLE 1 RT means and accuracy rates, per research design

	Within-subject		Between-subject (2017)	
	S-CIT	E-CIT	S-CIT	E-CIT
RTs (ms)				
Probe	413 ± 50	504 ± 41	431 ± 44	523 ± 47
Irrelevant	405 ± 48	475 ± 39	419 ± 39	482 ± 40
Target	499 ± 47	561 ± 36	518 ± 43	574 ± 38
Filler-F		529.2 ± 46.7		526.8 ± 48.4
Filler-U		594.4 ± 38.7		615.0 ± 34.1
P–I	7.8 ± 16.2	29.3 ± 20.8	11.8 ± 19.8	40.7 ± 25.3
SMD_{P-I}	0.48 [0.22, 0.75]	1.41 [1.05, 1.76]	0.60 [0.47, 0.72]	1.61 [1.39, 1.83]
ARs (%)				
Probe	98.9 ± 1.8	97.2 ± 3.2	96.6 ± 3.7	94.2 ± 5.6
Irrelevant	99.4 ± 0.8	99.1 ± 1.0	97.7 ± 2.1	97.9 ± 2.4
Target	86.1 ± 7.2	80.8 ± 9.8	77.4 ± 9.6	76.3 ± 10.5
Filler-F		93.8 ± 4.6		93.2 ± 5.2
Filler-U		77.7 ± 10.7		71.1 ± 11.2
P–I	−0.46 ± 1.76	−1.90 ± 3.03	−1.04 ± 3.04	−3.70 ± 5.48
SMD_{P-I}	−0.26 [−0.51, 0.00]	−0.63 [−0.90, −0.35]	−0.34 [−0.46, −0.23]	−0.68 [−0.84, −0.51]

Abbreviation: RT, reaction time.

this difference—although this small interaction effect could also very well be just accidental.

The same ANOVA was performed for ARs (Table 1, Figure 1 right panel). The between-subject design had somewhat higher probe-irrelevant differences in general, $F(1,605) = 8.77$, $p = .003$, $\eta_p^2 = .014$, 90% CI [.003, .034]. More importantly, the interaction was not significant, $F(1,605) = 2.46$, $p = .118$, $\eta_p^2 = .004$, 90% CI [0, .017].

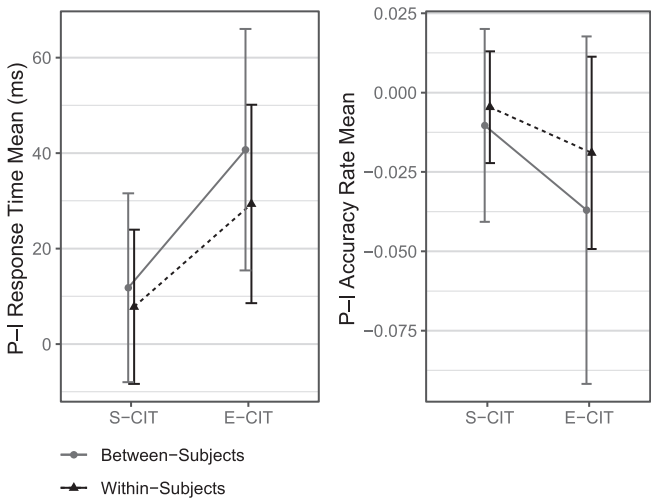

FIGURE 1 Means and*SD*s of individual probe-minus-irrelevant (P–I) response time mean differences (left panel) and accuracy rate differences (right panel); per version (S-CIT and E-CIT), and per research design (the present within-subject design versus the original between-subject design in the study by Lukács et al., 2017)

3.1 | Exploratory analysis: Order effects

To examine whether perhaps the order of tests in the present within-subject experiment influenced the results, an ANOVA was performed with the factors version (S-CIT vs. E-CIT) and test phase (first test vs. second test), on probe-irrelevant RT mean differences and on probe-irrelevant accuracy rate differences. The related aggregated values are displayed in Table 2.

For RTs, the version main effect was again significant, $F(1,59) = 69.11$, $p < .001$, $\eta_p^2 = .539$, 90% CI [.388, .637], but neither the test phase main effect nor the interaction was significant, $F(1,59) = 0.08$, $p = .782$, $\eta_p^2 = .001$, 90% CI [0, .047]; $F(1,59) = 0.49$, $p = .487$, $\eta_p^2 = .008$, 90% CI [0, .082].

Similarly for ARs, the version main effect was again significant, $F(1,59) = 0.97$, $p = .329$, $\eta_p^2 = .016$, 90% CI [0, .102], but neither the test phase main effect nor the interaction was significant, $F(1,59) = 12.12$, $p = .001$, $\eta_p^2 = .170$, 90% CI [.047, .307]; $F(1,59) = 2.19$, $p = .144$, $\eta_p^2 = .036$, 90% CI [0, .139].

4 | DISCUSSION

The practical implication of the larger probe-irrelevant differences in the E-CIT, as opposed to S-CIT, is straightforward: Adding familiarity-related fillers to the RT-CIT helps to more reliably reveal whether a person is concealing knowledge about a given critical information. While the results of the original paper may have been biased to some extent, the current within-subject replication shows that there is a

TABLE 2 RT means and accuracy rates, per test phase

	First test		Second test	
	S-CIT	E-CIT	S-CIT	E-CIT
RTs (ms)				
Probe	441 ± 43	502 ± 47	387 ± 42	506 ± 35
Irrelevant	433 ± 45	471 ± 40	380 ± 36	479 ± 39
Target	524 ± 46	546 ± 34	476 ± 35	577 ± 32
Filler-F		520.6 ± 49.3		538.8 ± 42.4
Filler-U		580.7 ± 39.0		609.5 ± 32.8
P–I	8.2 ± 17.0	30.7 ± 20.2	7.5 ± 15.6	27.8 ± 21.7
SMD_{P-I}	0.48 [0.09, 0.86]	1.52 [1.00, 2.03]	0.48 [0.11, 0.84]	1.28 [0.78, 1.77]
ARs (%)				
Probe	98.1 ± 2.1	97.2 ± 3.5	99.6 ± 1.0	97.3 ± 2.9
Irrelevant	99.2 ± 0.9	99.2 ± 0.7	99.6 ± 0.5	99.1 ± 1.4
Target	87.3 ± 6.9	81.1 ± 9.7	85.1 ± 7.4	80.4 ± 10.1
Filler-F		94.1 ± 4.5		93.5 ± 4.8
Filler-U		79.2 ± 10.7		76.1 ± 10.7
P–I	−1.03 ± 2.14	−1.97 ± 3.29	0.06 ± 1.13	−1.82 ± 2.76
SMD_{P-I}	−0.48 [−0.86, −0.09]	−0.60 [−0.97, −0.22]	0.05 [−0.30, 0.40]	−0.66 [−1.06, −0.25]

Note: Means and *SD*s (in the format of *M* ± *SD*) for individual RT means (RTs), and accuracy rates (ARs); per version (S-CIT and E-CIT), and per test phase (first test and second test).
Abbreviation: RT, reaction time.

large difference even when selective attrition is prevented. Thus, it is now assured that the E-CIT outperforms the S-CIT.

While the enhancing effect of the filler has now been repeatedly demonstrated, the underlying mechanisms are yet to be explored. The idea that responding is easier (and thus faster) when closely related items share the same response key, and vice versa, is a well established mechanism supported by dozens of studies (in particular in relation to the Implicit Association Test e.g., Greenwald et al., 2009; Nosek, Greenwald, & Banaji, 2007; but see also e.g., Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976, p. 435; Iordan, Greene, Beck, & Fei-Fei, 2015).

The increased task complexity implies increased task difficulty, and hence increased cognitive load. This could affect the task in different ways, and disentangling how this may influence the outcomes is not straightforward. As explained in the introduction, one plausible benefit is that increased cognitive load induces more attention to the presented items and their meaning. In any case, the notion that increased cognitive load facilitates slower deceptive responses is supported by previous RT-CIT papers (Hu, Evans, Wu, Lee, & Fu, 2013; Visu-Petra et al., 2013; Visu-Petra, Miclea, & Visu-Petra,- 2012), and even by studies of other deception detection methods (e.g., Vrij et al., 2008).

Related to task difficulty, another aspect that is yet to be investigated is whether or not the E-CIT is susceptible to countermeasures, as it has been shown for other RT-based deception detection methods (e.g., Hu, Chen, & Fu, 2012; Hu, Rosenfeld, & Bodenhausen, 2012; Van Bockstaele et al., 2012; Verschuere, Prati, & De Houwer, 2009). In particular, examinees might be able to manipulate probe-irrelevant RT differences by making deliberately fast responses to probes and/or slow responses to irrelevants. It seems possible that due to the relatively complex design of the E-CIT, examinees may find it difficult to intentionally alter the timing of their responses to only the appropriate subset of the items used (e.g. increase the response times only for the irrelevants by inhibiting fast reactions to these items, and yet still make fast responses to probes). Nonetheless, ultimately, this is an empirical question.

4.1 | Disadvantages and advantages of the online setting

A laboratory-based replication of the E-CIT is still needed. Even when preventing selective attrition, the online setting may allow more noise in the data than strictly controlled lab experiments due to (a) potentially sub-optimal computer hardware used by participants, (b) varying environmental factors (e.g., different lighting conditions or disruptive noises), and (c) participants' lower motivation to pay close attention and perform the task properly. Still, there are numerous studies that have explored the reliability of online psychological research (e.g., Buhrmester, Kwang, & Gosling, 2011; Paolacci & Chandler, 2014), in particular the validity of online RT research (e.g., Germine et al., 2012; Hilbig, 2016; McGraw, Tew, & Williams, 2000; for the HTML5/JavaScript framework as in the present study, see Reimers & Stewart, 2015), and recently even specifically the validity of the RT-CIT in online settings (Kleinberg & Verschuere, 2015; Verschuere & Kleinberg, 2015)—all of which unanimously conclude that online RT research such as in the present study is a sound alternative to conventional lab studies and that results obtained in this environment closely reflect those obtained in strictly controlled laboratory conditions. Note also that, in all RT-CIT experiments, the key dependent variable (probe-irrelevant difference) is obtained, for each participant, via a within-subject comparison, which also serves as a control for external influences. However, even more assurance is gained by using an overall within-subject research design (i.e., within-subject statistical comparisons), as in the present study.

Importantly, online research has its own advantages as well. Fast and relatively inexpensive data collection allows for hundreds of participants to be collected within days. The available population is very diverse and even international, thus it can provide a broad demonstration of generalizability, and the obtained samples also more closely reflect the test results of possible criminal suspects than a study that involves only university students as in typical lab studies. Furthermore, the RT-CIT itself may be applied using online (web browser-based) applications in real life cases (Lukács, 2019; Verschuere & Kleinberg, 2015): In any scenario where other appropriate software has not been set up, the RT-CIT can be easily administered via an online application using any web browser on a standard computer (or even on smartphones; Lukács, Kleinberg, Kunzi, & Ansorge, 2020).

Nonetheless, future related studies involving online participants should preferably avoid between-subject designs to prevent potential confounds arising from selective attrition – in particular when different conditions involve different levels of difficulty in performing the task.

4.2 | Effect of test order

Almost all previous RT-CIT studies so far used between-subject design, and none has examined the effect of repeated testing. The results of the present study indicate that repeated testing (hence familiarity with the task, practice, fatigue, etc.) has no substantial impact on the key outcomes (i.e., on the probe-irrelevant differences). The aggregated means of probe-irrelevant differences ("P–I") in Table 2 seem to be in line with the lack of significant test phase (i.e., test order) effects. The RT measure in particular seems unaffected (including consistently higher values for E-CIT). Nominally (i.e. despite no statistically significant difference), the accuracy rate differences seem to diminish in the S-CIT in the second phase. Nonetheless, the differences between probes and irrelevants are still nominally larger for E-CIT in both test phases.[4] In any case, ARs are never used as primary predictors in the RT-CIT (as indicated by its very name), and are generally only reported as a convention (and to rule out speed-accuracy tradeoff).

All in all, this invites future RT-CIT studies to use within-subject research design. This not only prevents confounds due to selective attrition (as long as the design is properly counter-balanced), but it is

in general more resourceful as well as more reliable than the between-subject design (as long as the order effect does not confound the results).

However, since the present study used two different CIT versions per subject, it does not provide sufficient evidence for the practical question of whether repeated testing of the same version may influence outcomes (such as deliberate practice by a suspect who is aware of an upcoming CIT test). This issue would require its own dedicated study.

5 | CONCLUSIONS

The enhancement of the RT-CIT with filler items may be a great advantage with important practical implications (Lukács et al., 2017). However, a subsequent study demonstrated robust differences in participant dropout rates between the key conditions (60–63% vs. 18–19%; Lukács & Ansorge, 2019). For a proper appraisal of the enhancement, the present study compared the two conditions in a within-subject design, to avoid differences in dropout rates, and ascertained that the difference is real and not due to a confound.

ACKNOWLEDGMENTS

Many thanks to Claudia "Kiki" Kawai for thoroughly proofreading the manuscript. Gáspár Lukács is a recipient of a DOC Fellowship of the Austrian Academy of Sciences at the Department of Cognition, Emotion, and Methods in Psychology at the University of Vienna. This funding source had no role or involvement related to the present study.

CONFLICT OF INTEREST

The author has no financial or personal conflicts of interest.

DATA AVAILABILITY STATEMENT

All data collected in the present study is available at https://osf.io/wu5cf. The data collected in the study of Lukács et al. (2017) is available at https://osf.io/kv65n/.

ORCID

Gáspár Lukács https://orcid.org/0000-0001-9401-4830

ENDNOTES

[1] Except for no probe recall check and hence no related exclusions—which is particular to the study of Lukács and Ansorge (2019).

[2] No participant had to be excluded due to too low accuracy; see preregistration for criteria.

[3] The items on these lists never contained any of the probes (i.e., the actual identity details of a given participant), but, within each category, they had the closest possible character length to the given probe (depending on the list of available items), and none of them started with the same letter (except in case of months). In case of countries, if the probe included a space (e.g., "New Zealand" or "Czech Republic"), the items on this list were all chosen to include a space as well.

[4] To note, larger probe-irrelevant differences in case of accuracy rates means lower negative values.

REFERENCES

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical Turk: A new source of inexpensive, yet high-quality data? *Perspectives on Psychological Science*, *6*(1), 3–5. https://doi.org/10.1177/1745691610393980

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use Welch's t-test instead of Student's t-test. *International Review of Social Psychology*, *30*(1), 92. https://doi.org/10.5334/irsp.82

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012). Is the web as good as the lab? Comparable performance from web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review*, *19*(5), 847–857. https://doi.org/10.3758/s13423-012-0296-9

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

Hilbig, B. E. (2016). Reaction time effects in lab- versus web-based research: Experimental evidence. *Behavior Research Methods*, *48*(4), 1718–1724. https://doi.org/10.3758/s13428-015-0678-9

Hu, X., Chen, H., & Fu, G. (2012). A repeated lie becomes a truth? The effect of intentional control and training on deception. *Frontiers in Psychology*, *3*, 1–7. https://doi.org/10.3389/fpsyg.2012.00488.

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

Hu, X., Rosenfeld, J. P., & Bodenhausen, G. V. (2012). Combating automatic autobiographical associations: The effect of instruction and training in strategically concealing information in the autobiographical implicit association test. *Psychological Science*, *23*(10), 1079–1085. https://doi.org/10.1177/0956797612443834

Iordan, M. C., Greene, M. R., Beck, D. M., & Fei-Fei, L. (2015). Basic level category structure emerges gradually across human ventral visual cortex. *Journal of Cognitive Neuroscience*, *27*(7), 1427–1446. https://doi.org/10.1162/jocn_a_00790

Kelley, K. (2018). MBESS: The MBESS R Package. R package version 4.4.3. Retrieved from 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. https://doi.org/10.1371/journal.pone.0118715

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, 863. https://doi.org/10.3389/fpsyg.2013.00863

Lawrence, M. A. (2016). Ez: Easy analysis and visualization of factorial experiments. R package version 4.4–0. Retrieved from https://CRAN.R-project.org/package=ez

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods*, *44*, 325–243. https://doi.org/10.3758/s13428-011-0146-0

Lukács, G. (2019). CITapp-a response time-based concealed information test lie detector web application. *Journal of Open Source Software*, *4*(34), 1179. https://doi.org/10.21105/joss.01179

Lukács, G., & Ansorge, U. (2019). Information leakage in the response time-based concealed information test. *Applied Cognitive Psychology*, *33*(6), 1178–1196. https://doi.org/10.1002/acp.3565

Lukács, G., Grządziel, A., Kempkes, M., & Ansorge, U. (2019). Item roles explored in a modified P300-based CTP concealed information test. *Applied Psychophysiology and Biofeedback*, *44*(3), 195–209. https://doi.org/10.1007/s10484-019-09430-6

Lukács, G., Kleinberg, B., Kunzi, M., & Ansorge, U. (2020). Response time concealed information test on smartphones. *Collabra: Psychology*, *6*(1), 4. https://doi.org/10.1525/collabra.255

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

McGraw, K. O., Tew, M. D., & Williams, J. E. (2000). The integrity of web-delivered experiments: Can you trust the data? *Psychological Science*, *11*(6), 502–506. https://doi.org/10.1111/1467-9280.00296

National Research Council. (2003). *Polygraph and lie detection*. Washington, DC: The National Academies Press. Retrieved from http://www.nap.edu/openbook.php?record_id=10420

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The implicit association test at age 7: A methodological and conceptual review. In J. A. Bargh (Ed.), *Social psychology and the unconscious: The automaticity of higher mental processes* (pp. 265–292). New York, NY: Psychology Press.

Olson, J., Rosenfeld, J. P., & Perrault, E. (2019). Deleterious effects of probe-related versus irrelevant targets on the "CIT effect" in the P300- and RT-based three-stimulus protocol for detection of concealed information. *Psychophysiology*, *56*(12), e13459. https://doi.org/10.1111/psyp.13459

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding mechanical Turk as a participant pool. *Current Directions in Psychological Science*, *23*(3), 184–188. https://doi.org/10.1177/0963721414531598

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

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

Reimers, S., & Stewart, N. (2015). Presentation and response timing accuracy in adobe flash and HTML5/JavaScript web experiments. *Behavior Research Methods*, *47*(2), 309–327. https://doi.org/10.3758/s13428-014-0471-1

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382–439. https://doi.org/10.1016/0010-0285(76)90013-X

Steiger, J. H. (2004). Beyond the F test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods*, *9*(2), 164–182. https://doi.org/10.1037/1082-989X.9.2.164

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

Van Bockstaele, B., Verschuere, B., Moens, T., Suchotzki, K., Debey, E., & Spruyt, A. (2012). Learning to lie: Effects of practice on the cognitive cost of lying. *Frontiers in Psychology*, *3*, 1–8. https://doi.org/10.3389/fpsyg.2012.00526

Verschuere, B., & Kleinberg, B. (2015). ID-check: Online concealed information test reveals true identity. *Journal of Forensic Sciences*, *61*(S1), S237–S240. https://doi.org/10.1111/1556-4029.12960

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., Prati, V., & De Houwer, J. (2009). Cheating the lie detector: Faking in the autobiographical implicit association test. *Psychological Science*, *20*(4), 410–413. https://doi.org/10.1111/j.1467-9280.2009.02308.x

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

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*, 1–11. https://doi.org/10.3389/fpsyg.2013.00146

Vrij, A., Mann, S. A., Fisher, R. P., Leal, S., Milne, R., & Bull, R. (2008). Increasing cognitive load to facilitate lie detection: The benefit of recalling an event in reverse order. *Law and Human Behavior*, *32*(3), 253–265. https://doi.org/10.1007/s10979-007-9103-y

Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology*, *111*(4), 493–504. https://doi.org/10.1037/pspa0000056

How to cite this article: Lukács G. Addressing selective attrition in the enhanced response time-based concealed information test: A within-subject replication. *Appl Cognit Psychol*. 2021;35:243–250. https://doi.org/10.1002/acp.3759