AMERICAN PSYCHOLOGICAL ASSOCIATION
SARMAC
Journal of Applied Research in Memory and Cognition

ISSN: 2211-3681
2022, Vol. 11, No. 2, 188–199
https://doi.org/10.1016/j.jarmac.2021.08.008

EMPIRICAL ARTICLE

Prolonged Response Time Concealed Information Test Decreases Probe-Control Differences but Increases Classification Accuracy

Gáspár Lukács
Department of Cognition, Emotion, and Methods in Psychology, University of Vienna

The Response Time Concealed Information Test (RT-CIT) can reveal that a person recognizes a relevant item (*probe*, e.g., a murder weapon) among other irrelevant items (*controls*), based on slower responses to the probe compared to the controls. The present paper assesses the influence of test length (due to practice, habituation, or fatigue) on two key variables in the RT-CIT: (a) probe-control differences and (b) classification accuracy, through a meta-analysis (using 12 previous experiments), as well as with two new experiments. It is consistently demonstrated that increased test length decreases probe-control differences but increases classification accuracies. The main implication for real-life application is that using altogether at least around 600 trials is optimal for the RT-CIT.

General Audience Summary

The Concealed Information Test (CIT) may be used to assess whether a person recognizes certain incriminating information details, for example, the murder weapon in a recent crime. One prominent type of CIT is the response time (RT)-based CIT, in which the critical details are repeatedly presented in a computerized task along with other irrelevant details (e.g., several weapons that were not used in the recent crime). Those who recognize the critical details (because, e.g., they participated in a crime) tend to respond slower to these details as compared to the irrelevant details, and based on the RT difference between critical and irrelevant details, they can be distinguished from those who do not recognize the critical details. In the present study, we show that when the details are presented many times (e.g., over 40 times each), results become more reliable. This means that with longer testing, more tested persons can be correctly distinguished as having or not having recognized the critical details.

Keywords: deception, concealed information test, response time, practice, habituation, test length

The Response Time Concealed Information Test (RT-CIT) is a computerized deception detection method that aims to disclose whether the tested person recognizes certain relevant items (*probes*), such as a weapon used in a recent homicide, among a set of other items (*controls*, a.k.a. "irrelevants"; Meijer et al., 2016; Suchotzki et al., 2017; Varga et al., 2014). During the RT-CIT, examinees have to categorize items that are presented on a computer screen by pressing one of two keys (e.g., either "E" or "I" on a regular keyboard). They are asked to press one of those keys (e.g., "E") when they see the probe or one of four controls, and they are asked to press the other key (e.g., "I") when they see a certain *target* (an additional control item designated for this purpose). When "guilty" examinees recognize the probe as the relevant item in respect of the deception detection scenario, they tend to have slower responses to it as compared to controls, and thereby they can be distinguished from "innocent" ones.[1]

Verschuere & De Houwer, 2011 remark that the RT-CIT may have an advantage over certain other types of CITs that use physiological indices because the latter is susceptible to habituation. However, whether or not the increased number of item presentations affect the RT-CIT has never actually been tested. The aim of the present paper is to assess, in the RT-CIT, the potential effects of test length on probe-control RT differences and classification accuracy.

This article was published Online First November 5, 2021.

Many thanks to Anna Walker and to Claudia "Kiki" Kawai for proofreading.

Gáspár Lukács has been funded by the OeAW Post-DocTrack Program. This funding source had no role or involvement related to this study.

The author declares that there is no conflict of interest.

Correspondence concerning this article should be addressed to Gáspár Lukács, Faculty of Psychology, University of Vienna, Vienna, Austria. Email: gaspar.lukacs@univie.ac.at

[1] The conditions in CIT studies are often labelled "knowledgeable" (here: "guilty") and "naive" (here: "innocent"), because in general this method does not assess lying per se, but the recognition of a certain detail. In the present paper, the illustrative designations "guilty" and "innocent" are used.

Test Length Effect on Probe-Control Differences

Probe-control (or "lie-truth") differences based on physiological measures such as electrodermal activity may indeed decrease when items are presented repeatedly (e.g., Thompson & Amy, 2009; or for CIT in specific, see Elaad & Ben-Shakhar, 1997; Lieblich et al., 1974). However, certain RT-based deception detection tests with "truth" versus "lie" (or "control") responses have also shown tendencies of diminished truth-lie effects with practice—although these findings are mixed and probably to a great extent depend on specific task designs and paradigms (Hu et al., 2012; Johnson et al., 2005; Van Bockstaele et al., 2012; Vendemia et al., 2005). In the RT-CIT specifically, slower responding to probes as compared to controls are generally assumed to originate in the perception of probes as response-incompatible stimuli (e.g., Lukács & Ansorge, 2021; Seymour & Schumacher, 2009; Verschuere & De Houwer, 2011).[2] Therefore, it is suggestive that various classic psychological RT tests that involve incompatible stimuli, to which comparatively slow responses are made, have also repeatedly been shown to be affected by practice, with RT differences between compatible and incompatible stimuli[3] decreasing with larger numbers of trials (e.g., Chen et al., 2013; D'Ascenzo et al., 2021; Gillebaart et al., 2020; Stroop, 1935).

Furthermore, from another perspective, a series of studies shows that more cognitively demanding RT-CIT designs (e.g., increased number of different items in the task) increase probe-control differences (Hu et al., 2013; Lukács & Ansorge, 2021; Verschuere et al., 2015; Visu-Petra et al., 2013). This could also imply that if participants can perform the task more easily after practicing it for some time, the decreased cognitive demand may lead to decreased probe-control differences.

Probe-Control Differences Versus Classification Accuracy

Two previous papers visually presented changes in the RT-CIT results as a function of test length. Firstly, Kleinberg and Verschuere (2016, p. 49, Figure 2) depicted, in two experiments, what appears to be a fairly stable increase of diagnostic efficiency (as measured by areas under curves; AUCs[4]) with increasing number of trials added to the analysis (in blocks of 36 trials, up to the full test of 360 trials). Related block-wise probe-control differences were not reported. Secondly, Hsu et al. (2020, p. 9, Figure 5, and p. 13, Figure 9) depicted,[5] also in two experiments, decreased probe-control differences in the guilty group: The data from each individual RT-CIT was divided into three epochs (760 trials into 253, 253, and 254 trials), and probe-control differences appeared smaller in the second as compared to the first, and the third as compared to the second epoch. Related diagnostic efficiency was not available since no innocent group was tested.

Even if both of these apparent findings ([a] general increase of AUCs with more trials included and [b] decrease of guilty probe-control differences with passing number of trials) were statistically significant and reliable, they would not necessarily contradict each other. In fact, this is exactly what has been found in several studies on the CIT based on autonomic responses: The differences in the electrodermal responses between probes and controls decreased with more item repetitions, but the overall classification accuracy increased (Ben-Shakhar & Elaad, 2002; Elaad & Ben-Shakhar, 1997; Lieblich et al., 1974). This is because diagnostic efficiency depends not only on the group mean of the probe-control differences but also on the variability among individual probe-control differences: In case of less variability (i.e., more consistent probe-control differences among guilty examinees and/or among innocent examinees), examinees can be more accurately classified (Lukács & Specker, 2020). Even if probe-control differences decrease in the later phase of the test (thereby also lowering the overall average of probe-control differences), as long as there is still a certain degree of difference, this additional data may be useful in decreasing variability by canceling out random noise (e.g., Lord & Novick, 2008). To mention an intuitive example, presenting each item altogether only one or two times would hardly provide a reliable outcome, as each response may be biased by random factors, such as blinking before an item is presented or a disrupting noise, and so forth. The more times each item is presented, the more such accidental variations are minimized in their influence on the averaged RT results.

It is however an empirical question whether, in the RT-CIT, there remains a sufficiently large probe-control difference in the later phases of the test to provide incremental value in classification accuracy.

In the present paper, Study 1 reports a meta-analysis conducted on 12 datasets obtained in previous RT-CIT experiments in order to examine the effects of test length on probe-control differences and on AUCs. In two new experiments, Studies 2a and 2b replicate the main findings of Study 1.

---

[2] For example, the response to controls may be conceptualized as a compatible (factually correct) "no" response to the question "is this item recognized as a unique relevant item in the task?", while the response to targets may be conceptualized as a "yes" to the same question. since it is a unique relevant item requiring a different response from the rest. The required response to the probe, however, is "no," which is thus incompatible to the factually correct "yes" response to the same question since it is actually a unique relevant item in that it is the one to-be-concealed detail.

[3] For example, in the well-known Stroop task (Stroop, 1935), the required (verbal) response "red" is compatible with red-colored stimuli, while the required response "green" would be incompatible with the same red-colored stimuli. The Stroop effect refers to that the incompatible responses tend to be slower than the compatible responses.

[4] The AUC is a diagnostic efficiency measure for binary classification that is based on a comparison of two distributions of all predictor values (in the present case, the distributions among guilty and among innocent participants; e.g., Rice & Harris, 2005). Given the predictors (here: probe-control RT mean differences), one may set a cutoff for classifying individuals: In the RT-CIT, a person with a predictor value above the cutoff (e.g., 30 ms difference) is classified as guilty, and those below the cutoff are classified as innocent. A receiver operating characteristic (ROC) curve can be constructed by plotting the true positive rates (ratio of correctly classified guilty individuals) as a function of false positive rates (ratio of incorrectly classified innocent individuals) across all possible cutoff points. The AUC is the area under this ROC curve. The AUC can range from 0 to 1, where 0.5 means chance level classification, and 1 means flawless classification (i.e., all guilty and innocent classifications can be correctly made based on the given predictor variable, at a given cutoff point).

[5] Hsu et al. (2020) also conducted a two-way ANOVA on the factors epoch (first, second, and third portion of the trials) and trial type, with the three trial (or item) types (a) probe, (b) control, and (c) target—and the resulting significant interaction was attributed to decreased probe-control differences, even though this could have been statistically assessed only by a one-way ANOVA for probe-control differences, or, alternatively, by including, in the two-way ANOVA, only the item types (a) probe and (b) control, but not the target, thereby ruling out the very plausible alternative that the significant interaction was caused by the target items.

Study 1

Method

The following analysis was performed on a recently published extensive database (Lukács & Specker, 2020) with trial-level results from RT-CIT studies, including 12 different experimental designs with both guilty and innocent participants, from seven different papers (Geven et al., 2020; Kleinberg & Verschuere, 2015, 2016; Lukács et al., 2017; Noordraven & Verschuere, 2013; Verschuere et al., 2015; Verschuere & Kleinberg, 2015; for detailed database description, see Lukács & Specker, 2020, pp. 5-6). Exclusion criteria followed exactly the paper reporting the database (Lukács & Specker, 2020, p. 6). Data was excluded from participants with an accuracy rate no higher than 75% for the main items (probes and controls merged), or an accuracy rate no higher than 50% for (a) target items or (b) target-side fillers or (c) nontarget-side fillers. "Fillers" are additional items presented throughout the task, used in one of the included experimental designs (Lukács et al., 2017). Target-side fillers are probe-referring expressions (e.g., the words "familiar" or "mine," when the probe is an autobiographical detail) and have to be categorized with the same key as the target (and, thus, opposite to the probe and the controls). Nontarget-side fillers are control-referring expressions (e.g., the words "unfamiliar" or "irrelevant") and have to be categorized with the same key as the probe and the controls. Fillers do not serve any diagnostic purpose, but their inclusion increases probe-control differences, thereby also increasing diagnostic efficiency based on these differences (for details, see Lukács & Ansorge, 2021; Lukács et al., 2017).

For all further calculations, responses that were incorrect, too slow (above 800 ms), or below 150 ms, were excluded.

None of the tests in Study 1 were preregistered: However, while only the analysis that was first performed is reported here, several alternate ways of analysis were also tested (e.g., different exclusion criteria, different divisions of the tests, etc.), and none of these differences changed the outcomes in any meaningful way, thereby attesting to the statistical robustness of the findings.

All analysis was conducted in R (R Core Team, 2020; with extension packages by Bates et al., 2015; Lukács, 2021; Robin et al., 2011; Viechtbauer, 2010; Wickham, 2016).

Results

Trial-Level Analysis

Probe, control, and target RTs, probe-control differences, and AUCs, from all participants per each dataset, as a function of trial number, are shown in Figures 1–3. All figures show *cumulative* data: The first data point is a baseline of the 50 first trials (in each single individual CIT), and all subsequent data points were obtained by adding five more trials one by one (in each CIT) and recalculating the related values. The majority of datasets show a trend of decreasing guilty probe-control differences (Figure 2), but, at the same time, almost all datasets show a clear trend of increasing AUCs (Figure 3).

The decrease of probe-control differences was statistically assessed using linear mixed modeling (LMM; Revise as Bates et al., 2015) on raw trial-level response times, including stimulus type (probe vs. control) and trial number (divided by 100 for easier interpretation) as fixed factors (main effects and interaction), with the random effects of participant and dataset as factors (intercept and the slope of Stimulus Type, i.e., varying overall RT and varying probe-control difference).

The model's total explanatory power is moderate ($R^2 = 0.24$; fixed effects alone: $R^2 = 0.02$). The stimulus type coefficient reflects simply the overall probe-control differences, $B = 33.1$ ms, 95% CI [25.4, 40.8], while the trial number coefficient indicates the overall practice effect (i.e., generally decreasing RTs) per every 100 trial, $B = 11.7$ ms, 95% CI [11.0, 12.5] (CIs calculated using Wald approximation). More importantly, the coefficient for the Stimulus Type × Trial Number interaction indicates the decrease of probe-control differences per 100 trial, $B = 2.5$ ms, 95% CI [1.6, 3.4]. Comparing this full model to the one without the Stimulus Type × Trial Number interaction shows that the interaction is a significant contributor, $\chi^2(1) = 31.5$, $p < .001$.

The decrease of 2.5 ms per 100 trials (0.025 ms per trial) might be relatively small, but it is certainly not negligible: By 600 trials, the total decrease would be 15 ms, which is a substantial drop relative to the overall average of 31.5 ms that distinguishes guilty from innocent.

Block-Level Analysis

For statistical assessments of the test length effect on diagnostic efficiency, one straightforward approach is to divide each single test into two halves (using the maximum trial number divided by two as dividing point).

Firstly, individual probe-control RT differences, based on the first half of the test versus the second half of the test (from here onwards: *Test Phase* effect) are to be compared to each other, to confirm the decrease from the first to the second half. Simple *t*-tests already show the significant difference when using all predictors (all datasets merged) from the first halves versus from the second halves of the tests, 5.12 ms, 95% CI [3.06, 7.18] ($M \pm SD = 27.73 \pm 34.13$ vs. 22.61 ± 30.39), $t(1099) = 4.88$, $p < .001$, $d = 0.15$, 95% CI [0.09, 0.21]; or when using the grand means per dataset (Figure 4, top Panel A), 4.51, 95% CI [1.95, 7.06] ($M \pm SD = 32.44 \pm 13.99$ vs. 27.93 ± 14.29), $t(11) = 3.88$, $p = .003$, $d = 1.12$, 95% CI [0.38, 1.84] (the predictors' Pearson correlation between the first and second halves: $r(10) = .960$, 95% CI [0.859, 0.989], $p < .001$).

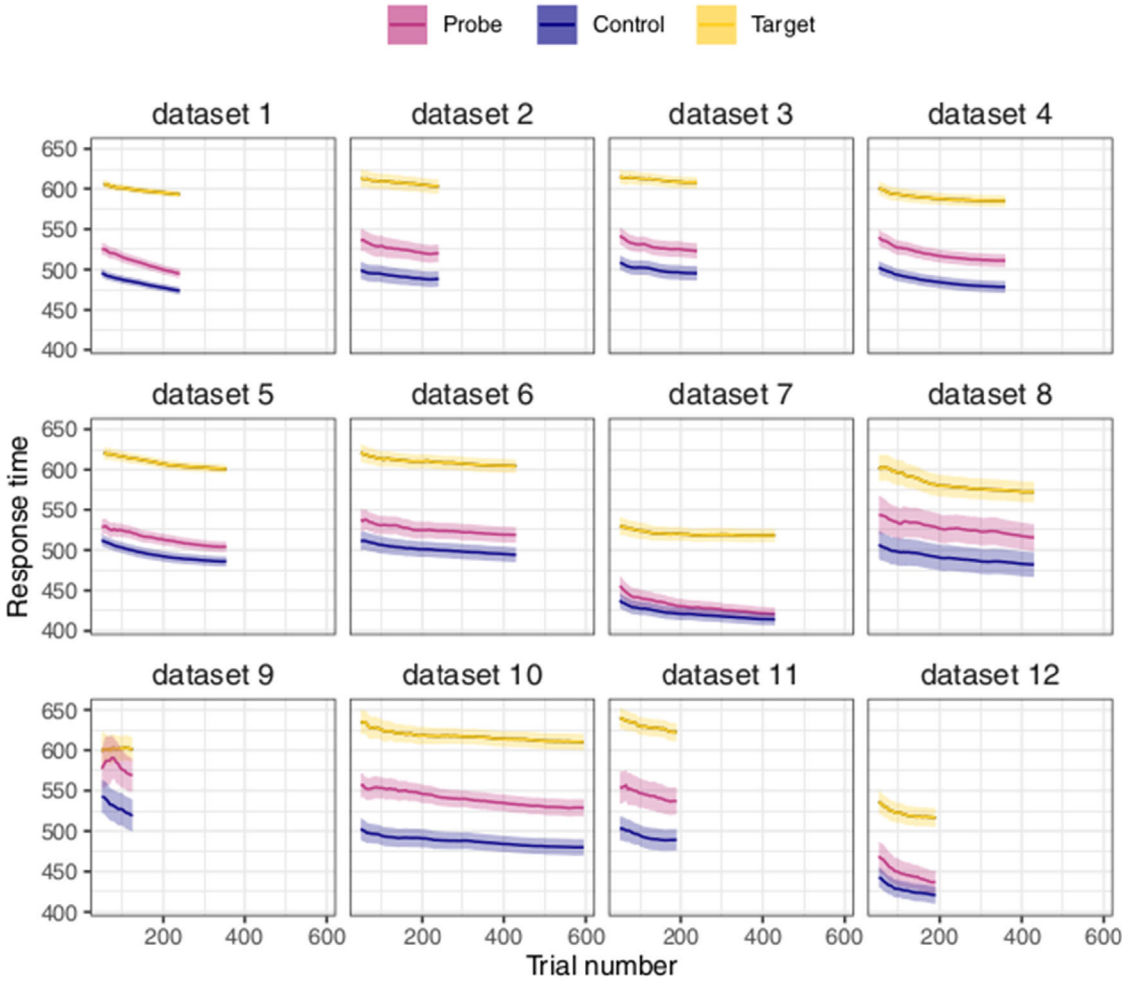
However, a meta-analytical comparison is performed as well, using the *rma* function of the *metafor* R package with default settings: This function calculates a random effects meta-analytic model as a special case of general linear model with known heteroscedastic sampling variances, using restricted maximum-likelihood estimator for standardized mean change, and weighting studies (here: "datasets") by inverse-variance (Viechtbauer, 2010).

The second half of the RT-CITs, as compared to the first half, was proven to yield smaller probe-control RT differences, with a standardized mean change of 0.133, 95% CI [0.053, 0.213], $p = .001$. The effect sizes per dataset depicted in Figure 5, and in the mean probe-control differences per dataset and Test Phase as depicted in the top Panel (A) of Figure 4, seem to indicate that the Test Phase effect (decrease in probe-control differences) is consistent across the datasets, in line with a nonsignificant heterogeneity test, $QE(11) = 12.70$, $p = .314$. Nonetheless, there is still a small amount of heterogeneity among the true effects contributing to the total variability in the effect size estimates, $I^2 = 27.7\%$, 95% CI [0, 63.4].

In contrast, the AUCs based on the full test were consistently larger than those based on the first half of the tests; see the bottom

Figure 1

Cumulative Means of Response Times Per Item Type



Note. Cumulative group means (and their 95% CIs) of individual RT means per item type, along with trial numbers (starting from trial 50), per each dataset (with numbering corresponding to that in Lukács & Specker, 2020; see also Figure 4), from guilty conditions only. See the online article for the color version of this figure.

Panel (B) of Figure 4. It is worth noting that there is only one exception where AUCs based on the full test are smaller, in dataset 7, but this can be explained by the fact that, in this specific dataset, the probe-control differences were very small in the first place (for the reasons why, see Verschuere et al., 2015), and by the second half of the tests they decreased on average to almost zero (see dataset 7 in Figure 2). Regardless, comparing the obtained AUCs in a *t*-test (first halves vs. full tests, paired by dataset), there is strong evidence for the overall AUC difference of 0.047, 95% CI [0.023, 0.070] ($M \pm SD$ = 0.796 ± 0.095 vs. 0.750 ± 0.090), $t(11) = 4.39$, $p = .001$, $d = 1.27$, 95% CI [0.48, 2.02]; Pearson correlation: $r(10) = .922$, 95% CI [0.740, 0.978], $p < .001$.

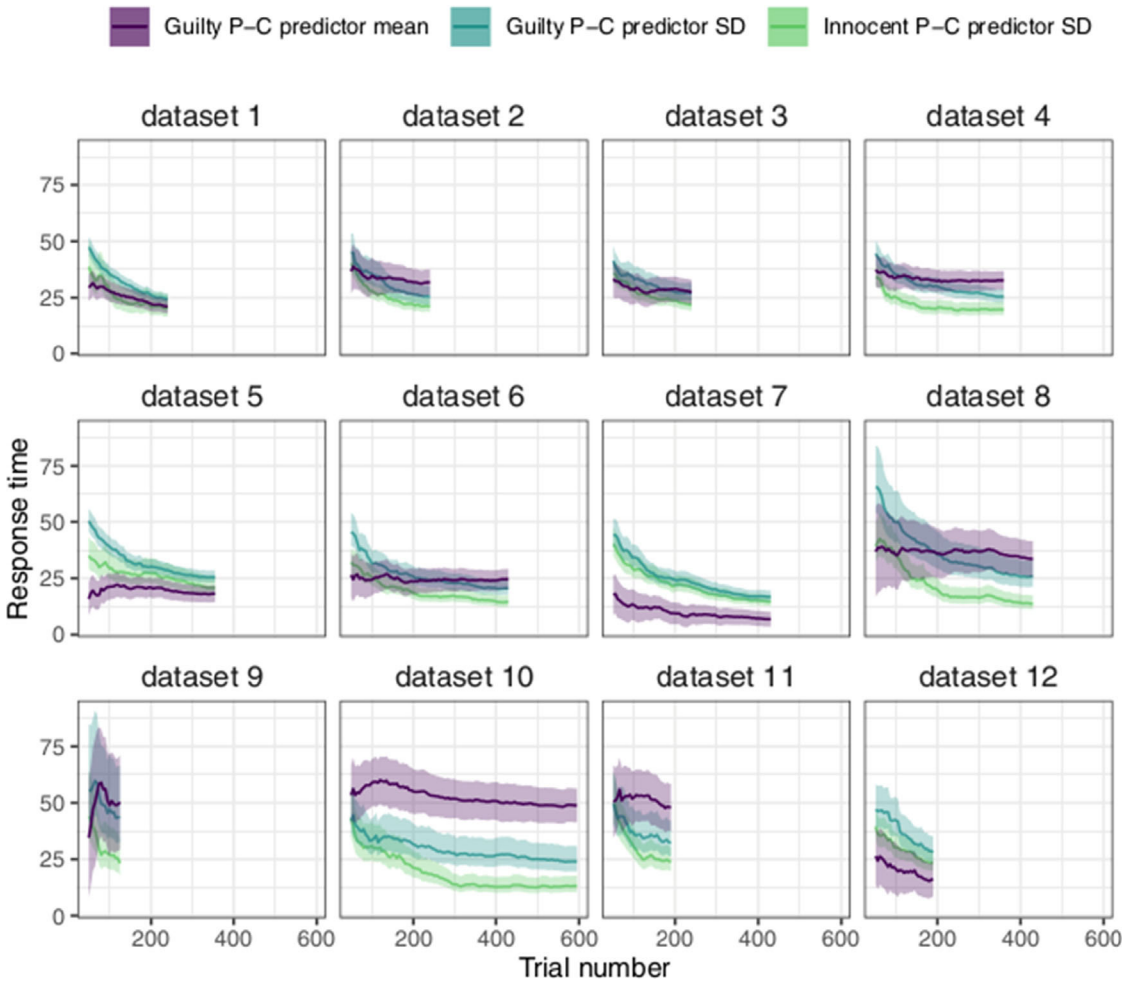
Comparing rates of correct classification (the average of true positive and true negative rates) gives similar results with a raw mean difference of 0.040, 95% CI [0.020, 0.059] ($M \pm SD = 0.758 \pm 0.081$ vs. 0.719 ± 0.082), $t(11) = 4.42$, $p = .001$, $d = 1.28$, 95% CI [0.49, 2.03]; Pearson correlation: $r(10) = .927$, 95% CI [0.755, 0.980], $p < .001$. A DeLong's test on the AUCs calculated using all predictors (all datasets merged) from the first halves versus on the full test further strongly confirms the findings with an AUC difference of 0.042, 95% CI [0.028, 0.057], $D = 5.52$, $p < .001$ (0.726, 95% CI [0.703, 0.749] vs. 0.769, 95% CI [0.747, 0.790]).

Study 2a and 2b

A recently conducted series of experiments introduced an RT-CIT variation that serves to detect concealed language knowledge (Lukács et al., in press): For example, a suspect, speaking in English, may claim to not understand anything in Polish. In that case, Polish words could serve as probes in the RT-CIT, while graphemically similar pseudowords would serve as controls. It was demonstrated that persons who understand the tested language recognize the real word and tend to have slower responses to it as compared to the pseudowords, and thereby they can be distinguished from those who do not understand the language. Relevant to the present paper, the same study also examined test length effects (unrelatedly to the idea

Figure 2

Cumulative Means and SDs of Probe-Control Response Time Differences



Note. Cumulative group means (guilty condition only) and SDs (guilty and innocent) of individual probe-control (*P-C*) RT differences (as predictors for potential diagnostics), including 95% CIs, along with trial numbers, per each dataset. As explained in the Introduction, while the decreasing probe-control RT differences are detrimental to classification accuracy, the correspondingly decreasing SDs are beneficial to it (cf. Figure 3). The SDs here are not to be confused with within-subject SDs. To clarify this using an example, the calculation for dataset 1, which includes 38 guilty participants, was as follows (regardless of the number of trials included). First, for each of the 38 participants, the mean of all valid control RTs was subtracted from the mean of all valid probe RTs. This resulted in 38 values of probe-control RT differences, which are the conventional predictors of the Response Time Concealed Information Test (RT-CIT). Subsequently, the one *mean* (with 95% CI) and the one *SD* (with 95% CI) of these 38 predictor values was calculated. Each data point in the figure depicts the latter mean and SD (per included trial number and per dataset). See the online article for the color version of this figure.

of detecting concealed language). The results are reported in the following.

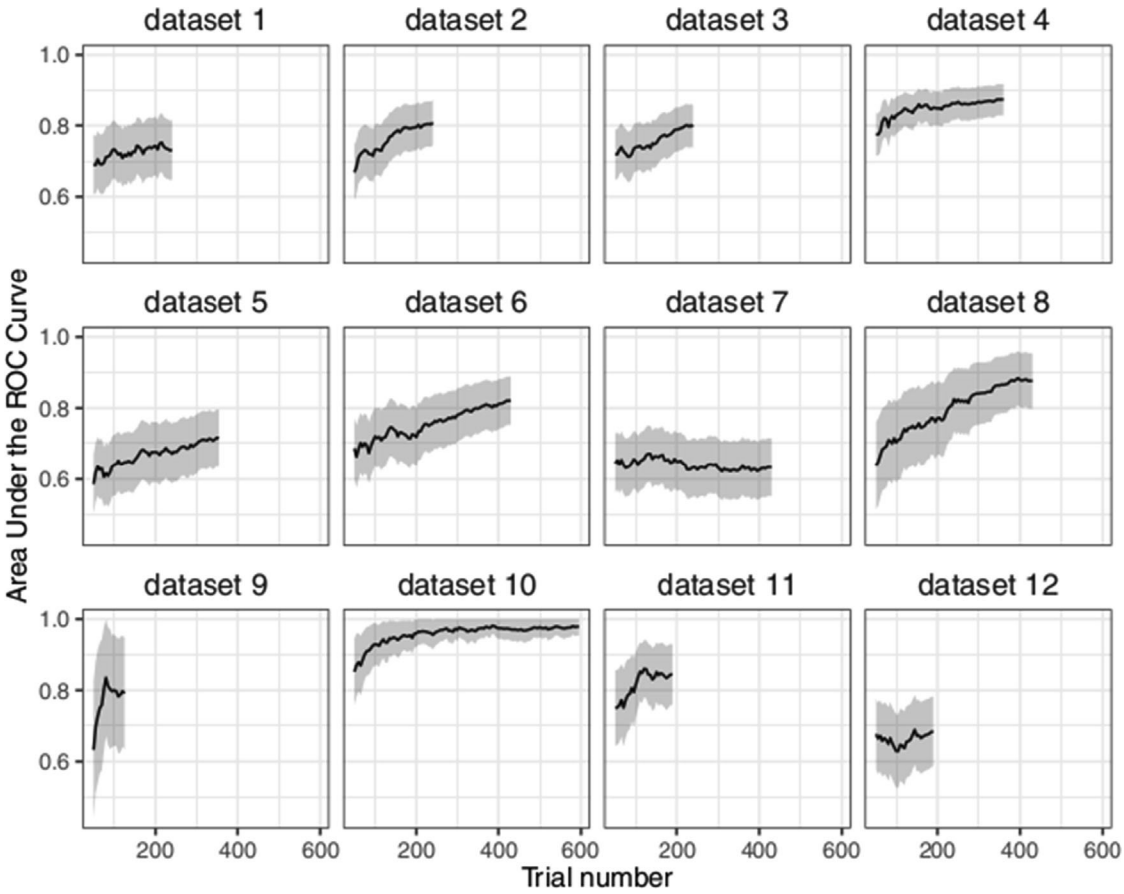
Method

For both studies, preregistered stopping rules and exclusion criteria were followed (Study 2a: https://osf.io/2g76c/; Study 2b: https://osf.io/gdk92/). The tests for block-wise comparisons of probe-control differences and AUCs were preregistered for Study 2a only, but for consistency and for further evidence the exact same tests are reported exploratorily for Study 2b as well. The LMM analysis was not preregistered.

Participants

In Study 2a, participants were 50 Polish and 50 Hungarian natives, fluent in English as a second language, recruited via the online crowdsourcing platform Prolific (https://www.prolific.co/). The (preregistered) sample size was based on the estimated available participants on Prolific. All participants were tested for Hungarian as well as Polish (simulating a scenario where two different concealed languages are suspected), hence serving as each other's control groups. Out of the 100, five had to be excluded (two for too low accuracy, three for not selecting correctly at least three out of the four probes in their native language in an attention check at the end

Figure 3
Cumulative Areas Under the Curves



Note. Cumulative AUCs, based on individual probe-control RT difference predictor values (guilty versus innocent), along with trial numbers, per each dataset.

of the test), leaving 49 Hungarian (age = 26.2 ± 6.8; 38 male) and 46 Polish speakers (age = 25.1 ± 6.9; 37 male). Participants were paid 4 .88 GBP for the 40–45 min experiment and a potential 0.50 GBP bonus for not having been detected in either language (hence altogether max. 1.00 GBP).[6]

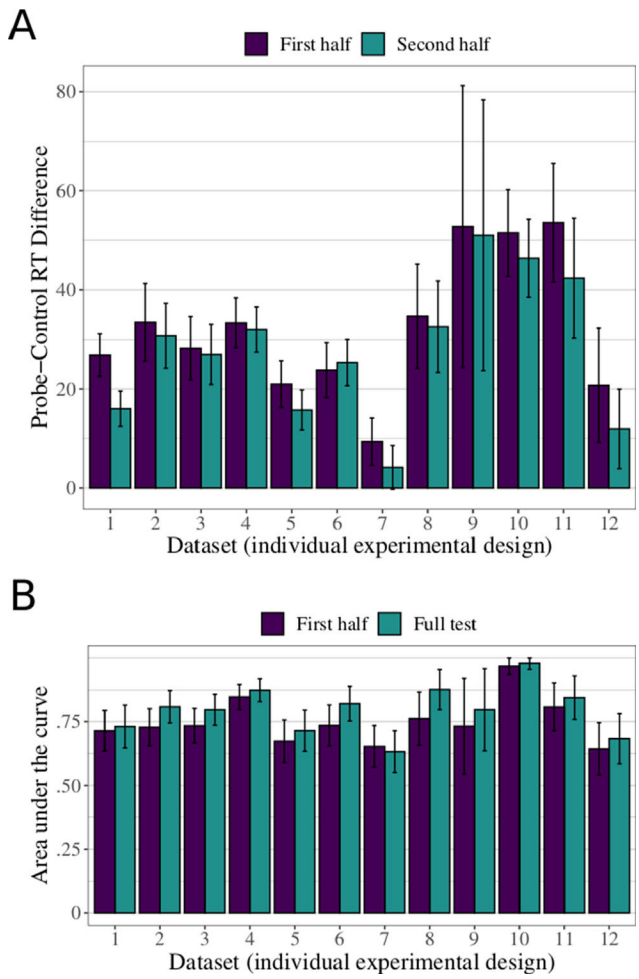
In Study 2b, participants (again recruited via Prolific) were 70 English monolinguals and 130 Russian native speakers fluent in English as a second language, and all were tested for knowledge of the Russian language. The sample size was decided based on optional stopping (for details, see the preregistration at https://osf.io/tkd74/). Out of these, eight had to be excluded (one for not selecting probes correctly, seven for too low accuracy), leaving 124 Russian natives (age = 31.8 ± 10.4 [one unknown]; 46 male, 78 female) and 68 English monolinguals (age = 31.1 ± 9.4; 36 male, 32 female). Participants were paid 3.28 GBP for the 25–30 min experiment, and a potential 0.50 GBP bonus for not having been detected as understanding Russian.

Procedure

All instructions were in English. Participants were asked to imagine themselves, during the testing, in a scenario where it would be crucial for them to conceal their knowledge of each given tested language (Polish, Hungarian, or Russian). Hence, here, speakers of the given language represent "guilty" examinees, while nonspeakers represent "innocent" ones. The main task for each tested language contained four blocks, each block with its own unique set of probe, target, and four controls. Thereby, for each participant, the task contained altogether eight blocks in Study 2a (two tests: for Polish and for Hungarian), while four blocks in Study 2b (one test: for Russian). The probes and targets were always real and meaningfulness-referring words in the tested language (Polish or Hungarian), while the controls were graphemically similar pseudowords (indistinguishable from real words for nonspeakers). In each block, each of these items was repeated 18 times, thereby one block contained altogether 108 items (18 probe, 18 target, and 72 controls). Additionally, there were three different target-side fillers (expressions referring to meaningfulness and genuineness; e.g., the English words "meaningful" or "true"), and six different nontarget-side fillers (expressions referring to meaninglessness

[6] Successful detection for this purpose and for automatic feedback, was based on a $d = 0.3$ (standardized mean difference) between probe and control RTs, a higher level than in previous studies to favor participants (Noordraven & Verschuere, 2013).

Figure 4
Meta-Data: Probe-Control Differences Versus Areas Under the Curves



Note. The average individual probe-control RT differences in guilty groups from the first and the second halves of the tests (Panel A), and AUCs based on the first half and on the full tests (Panel B). Error bars indicate 95% CIs in both cases. See the online article for the color version of this figure.

and fakeness; e.g., the English words "untrue" or "fake"). Each filler was repeated six times per block; hence, each block contained altogether 54 filler trials and a total number of 162 trials.

The order of the items was randomized in groups: first, all six items (one probe, four irrelevants, and one target) in the given block 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). Fillers were placed among these items in a random order, but with the restrictions that a filler trial was never followed by another filler trial, and each of the nine fillers preceded each of the main items (probe, target, and each of the four controls) exactly one time.

Participants had to press the key "I" whenever the target or target-side filler appeared, while they had to press the key "E" whenever the probe, a control, or a nontarget-side filler appeared. The inter-trial interval always randomly varied between 0.5 and 0.8 s. In case of an incorrect response or no response within 1 s, the caption "False!" or "Too slow!", respectively, appeared in red color for 0.5 s, followed by the next trial. There were three short initial practice rounds that included all items from the upcoming first block, and participants had to repeat any round on which they had too few correct responses in time. For analysis, only trials with a correct response between 0.15 s and 1 s were used.

Results

Trial-Level Analysis

The decrease of probe-control differences was again assessed using LMM on raw trial-level response times. Here, apart from stimulus type (probe vs. control) and trial number (divided by 100), block number was also included as a fixed factors. The two interactions of interest were also included: Stimulus Type × Trial Number and Stimulus Type × Block Number.[7] The random effects of participant as factor were also included (intercept and the slope of stimulus type).

The model's total explanatory power is substantial (R^2 = 0.33; fixed effects alone: R^2 = 0.11). As before, the stimulus type coefficient reflects the overall probe-control differences, B = 112.6 ms, 95% CI [104.5, 120.8], while the trial number coefficient indicates the overall practice effect per every 100 trial, B = –20.4 ms, 95% CI [–22.1, –18.8]. In contrast, the block number coefficient indicates that, although subsumed the larger trial-by-trial decrease, overall RTs to some degree increase per each new block (with new items in the task), B = 19.3 ms, 95% CI [16.5, 22.0].

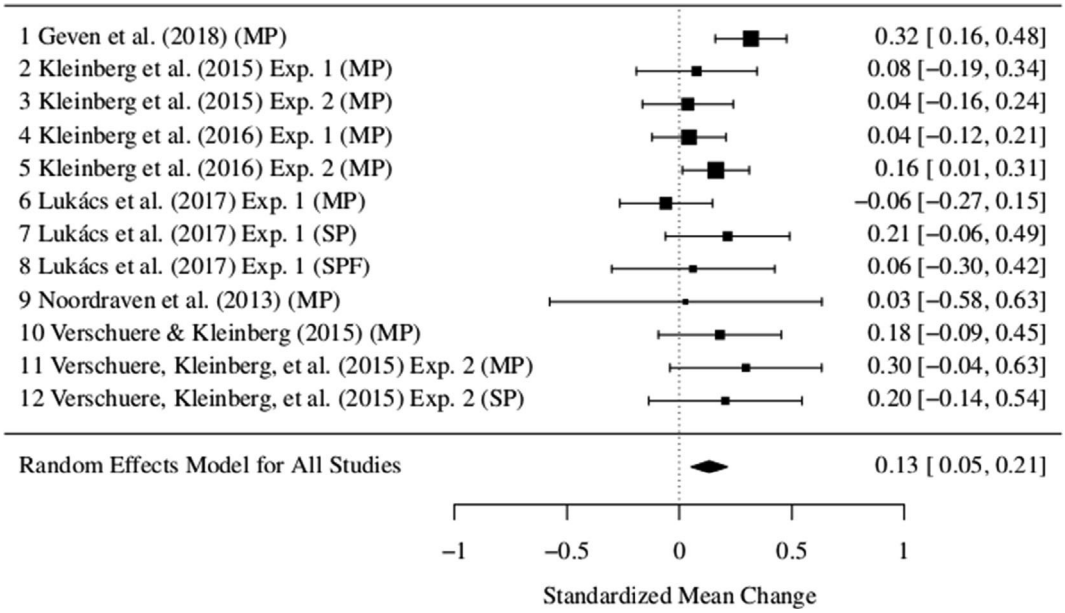
More importantly, the coefficient for the Stimulus Type × Trial Number interaction again indicates the decrease of probe-control differences per 100 trial, B = 5.45, 95% CI [1.7, 9.2]. Comparing the full model to the one without the Stimulus Type × Trial Number interaction shows that the interaction is a significant contributor, $\chi^2(1) = 8.0$, $p = .005$. At the same time, the coefficient for the Stimulus Type × Block Number interaction seem to indicate an additional decrease of probe-control differences per block, B = 4.35 ms, 95% CI [2.0, 10.7]. However, comparing the full model to the one without the Stimulus Type × Block Number interaction does not show this interaction to be a significant contributor, $\chi^2(1) = 1.8$, $p = .177$, indicating that the changing blocks do not substantially affect probe-control differences.

Block-Level Analysis

The one-way ANOVA on the probe-control RT differences (in guilty conditions only) with Block as a four-level within-subject factor (data from Blocks 1, 2, 3, and 4, each separately) showed, again, clear probe-control difference decrease with time in both experiments: $F(3, 282) = 10.45$, $p < .001$, $\eta_p^2 = .100$, 90% CI [0.045, 0.151], $\eta_G^2 = .055$ (Study 2a), and $F(3, 369) = 13.91$, $p < .001$, $\eta_p^2 = .102$, 90% CI [0.053, 0.147], $\eta_G^2 = .052$ (Study 2b); see Figure 6, top Panel (A). One-sided DeLong's tests however also showed, in both

[7] The Block Number × Trial Number and the three-way Stimulus Type × Block Number × Trial Number interactions are of no particular interest, and their inclusion would lead to an unnecessarily complex model, and, in any case, it makes no difference in the reported key results.

Figure 5
Test Phase Effect Estimates



Note. Numbers indicate the dataset number (corresponding to those in the other figures). SP= Single-probe protocol; MP= multiple-probe protocol; SPF= single-probe protocol with fillers (see Lukács et al., 2017; Verschuere et al., 2015).

cases, that AUCs were larger in case of using the data from the full test as compared to when using the data from the first block only by 0.031, 90% CI [0.005, 0.057] (D = 1.98, p = .024; Study 2a), and by 0.036, 90% CI [0.004, 0.068] (D = 1.83, p = .033; Study 2b), and also as compared to when using the data from the first two blocks by 0.020, 90% CI [0.003, 0.036] (D = 2.00, p = .023; Study 2a), and by 0.022, 90% CI [–0.001, 0.046] (D = 1.56, p = .059—this last just somewhat above alpha level; Study 2b); see Figure 6, bottom Panel (B).

General Discussion

That probe-control differences are diminishing as a function of test length could constitute a problem in the case of an excessively lengthy test. However, RT-CIT studies so far have never used more than 600 trials, and, up to this length, despite decreasing probe-control differences, classification accuracy increased steadily with the cumulative inclusion of trials up to the very end in almost all included datasets (see Figures 3 and A1). The straightforward implication for real-life use is that the RT-CIT should contain at least around 600 trials for optimal results.

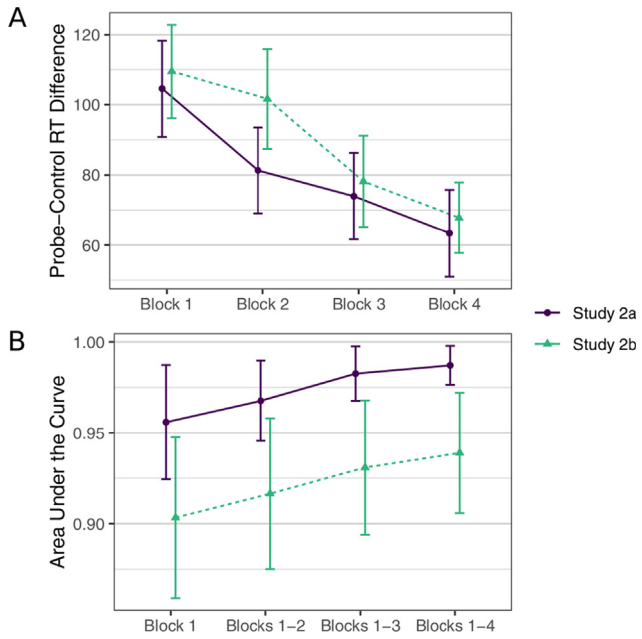
It is, however, unlikely that many thousands of trials would still have incremental diagnostic value. After a certain number of trials, practice, let alone fatigue, may completely diminish probe-control differences and actually reduce classification accuracy (see the Appendix). Relatedly, computer simulations of RT-CIT data that assume constancy in probe and control RTs throughout the task would lead to misleading outcomes (i.e., overly optimistic outlook on diagnostic improvements with more trials; see Koller et al., 2020, pp. 1412, 1415). Models would, in any case, require validation on real-life RT-CIT results, but if precise enough, they may be of great use in estimating not only optimal trial numbers but also overall classification accuracy as well as classification accuracy per probe (the latter via taking only the given relevant section of trials into account).

Another implication is that when probes are tested sequentially, probes that are tested later are liable to show smaller differences from controls. Therefore, the most crucial available probes (e.g., ones that are central in the investigated crime and that are certain not to be known to anyone except the guilty person) should be tested first. If several nonessential items (whose results would be less decisive in any case) are tested in the beginning, this may squander the most effective range of trials in the RT-CIT.

There is increasing evidence that the RT-CIT is resistant to countermeasures that aim to manipulate the RTs during the test, such as deliberate attempts at responding faster to probes (Norman et al., 2020; Suchotzki et al., 2021). However, based on the present findings, the RT-CIT may actually be vulnerable to countermeasure efforts via practicing the task in advance of official testing. The person about to be tested does not necessarily have to know the probe, control, and target items in the task for this: For example, in Studies 2a and 2b, each block contained different items, and yet probe-control differences decreased linearly throughout the task, regardless of the blocks (Figure A1). Even if the RT-CIT were not already freely available (Lukács, 2019), the essence of the task can be easily implemented in various free applications (e.g., Peirce & MacAskill, 2018). What is more, practice with similar target-nontarget classification tasks, which are also widely available online (e.g., Sochat, 2018), is probably at least to some extent transferable to the RT-CIT (see e.g., D'Ascenzo et al., 2021; Proctor & Lu, 1999; Tagliabue et al., 2000). Nonetheless, the test length effects in the present study, with RT-CIT lengths up to around 600 trials, were

Figure 6

Probe-Control Differences Versus Areas Under the Curves, Block-Wise, in Studies 2a and 2b



Note. The average individual probe-control RT differences in guilty groups in each separate block (Panel A), and AUCs based on the different numbers of blocks included (Block 1: first block only, Block 1–2: first two blocks, etc.; Panel B); from Study 2a and 2b. Reflecting the statistical findings, probe-control differences decrease with longer testing, while AUCs increase. Error bars indicate 95% CIs in both cases. See the online article for the color version of this figure.

generally small in magnitude (see Figures 2, 4, and A1). It would require further dedicated experiments to see whether the more extensive practice can substantially reduce classification accuracy in a subsequent RT-CIT.

Finally, the presented findings serve as yet another practical case in which increased classification accuracy cannot be inferred from increased probe-control differences, but the two, in fact, go in opposite directions as a function of the independent variable (test length; Lieblich et al., 1974). More generally, the demonstrated effects of lengthy testing should be carefully considered (in particular regarding meticulous counterbalancing) in the case of within-subject study designs using the RT-CIT in the future.

Combating Probe-Control Difference Decrease

If probe-control differences decrease because practice makes the task easier, and especially if this happens due to getting used to the items in the task, then participants continually having to process new items may re-elevate the cognitive demand of the task and re-enforce attention. One attempt at such a manipulation (fillers changing continually or block-wise) is described in the online Appendix A (https://osf.io/tkd74/). The findings are mixed, with some tentative but not unambiguous evidence in support of this proposal. Future attempts may aim at a more gradual and straightforward manipulation of task difficulty, for example, via an RT limit that is dynamically adjusted, for each new trial throughout the test, based on the examinee's performance (e.g., error rate or mean RT) during the latest section of the test (e.g., last 100 trials).

One could also attempt to statistically correct for the decreasing probe-control differences. Along with the probe-control difference decrease, the differences between the different control items may also decrease with a longer test. In that case, it may help to standardize probe-control differences, within each block, relative to the differences between control items (Elaad & Ben-Shakhar, 1997), for example, by dividing the probe-control mean RT difference with the standard deviation (*SD*) of the control items (Noordraven & Verschuere, 2013; Verschuere et al., 2015, p. 64). However, contrary to what has been found in the CIT based on autonomic responses (Elaad & Ben-Shakhar, 1997), there seems to be no benefit of within-block standardization for the RT-CIT data of the present studies (Studies 1, 2a, or 2b; for details, see the online Appendix B).

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Ben-Shakhar, G., & Elaad, E. (2002). Effects of questions' repetition and variation on the efficiency of the guilty knowledge test: A reexamination. *Journal of Applied Psychology*, *87*(5), 972–977. https://doi.org/10.1037/0021-9010.87.5.972.

Chen, Z., Lei, X., Ding, C., Li, H., & Chen, A. (2013). The neural mechanisms of semantic and response conflicts: An fMRI study of practice-related effects in the Stroop task. *NeuroImage*, *66*, 577–584. https://doi.org/10.1016/j.neuroimage.2012.10.028.

D'Ascenzo, S., Lugli, L., Nicoletti, R., & Umiltà, C. (2021). Practice effects vs. transfer effects in the Simon task. *Psychological Research*, *85*(5), 1955–1969. https://doi.org/10.1007/s00426-020-01386-1.

Elaad, E., & Ben-Shakhar, G. (1997). Effects of item repetitions and variations on the efficiency of the guilty knowledge test. *Psychophysiology*, *34*(5), 587–596. https://doi.org/10.1111/j.1469-8986.1997.tb01745.x.

Geven, L. M., Ben-Shakhar, G., Kindt, M., & Verschuere, B. (2020). Memory-based deception detection: Extending the cognitive signature of lying from instructed to self-initiated cheating. *Topics in Cognitive Science*, *12*(2), 608–631. https://doi.org/10.1111/tops.12353.

Gillebaart, M., Benjamins, J., van der Weiden, A., Ybema, J. F., & De Ridder, D. (2020). Practice makes perfect: Repeatedly dealing with response conflict facilitates its identification and speed of resolution. *Journal of Research in Personality*, *86*. https://doi.org/10.1016/j.jrp.2020.103955. Article 103955.

Hsu, A., Lo, Y.-H., Ke, S.-C., Lin, L., & Tseng, P. (2020). Variation of picture angles and its effect on the concealed information test. *Cognitive Research: Principles and Implications*, *5*(1). https://doi.org/10.1186/s41235-020-00233-6. Article 33.

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*. https://doi.org/10.3389/fpsyg.2012.00488. Article 488.

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.

Johnson, R., Barnhardt, J., & Zhu, J. (2005). Differential effects of practice on the executive processes used for truthful and deceptive responses: An event-related brain potential study. *Cognitive Brain Research*, *24* (3), 386–404. https://doi.org/10.1016/j.cogbrainres.2005.02.011.

Kleinberg, B., & Verschuere, B. (2015). Memory detection 2.0: The first web-based memory detection test. *PLOS ONE*, *10*(4). https://doi.org/10.1371/journal.pone.0118715. Article e0118715.

Kleinberg, B., & Verschuere, B. (2016). The role of motivation to avoid detection in reaction time-based concealed information detection. *Journal of Applied Research in Memory and Cognition*, *5*(1), 43–51. https://doi.org/10.1016/j.jarmac.2015.11.004.

Koller, D., Hofer, F., Grolig, T., Ghelfi, S., & Verschuere, B. (2020). What are you hiding? Initial validation of the reaction time-based searching concealed information test. *Applied Cognitive Psychology*, *34*(6), 1406–1418. https://doi.org/10.1002/acp.3717.

Lieblich, I., Naftali, G., Shmueli, J., & Kugelmass, S. (1974). Efficiency of GSR detection of information with repeated presentation of series of stimuli in two motivational states. *Journal of Applied Psychology*, *59*(1), 113–115. https://doi.org/10.1037/h0035781.

Lord, F. M., & Novick, M. R. (2008). *Statistical theories of mental test scores*. Information Age Publishing Inc.

Lukács, G. (2019). CITapp—A response time-based Concealed Information Test lie detector web application. *Journal of Open Source Software*, *4*(34). https://doi.org/10.21105/joss.01179. Article 1179.

Lukács, G. (2021). neatStats: An R package for a neat pipeline from raw data to reportable statistics in psychological science. *The Quantitative Methods for Psychology*, *17*(1), 7–23. https://doi.org/10.20982/tqmp.17.1.p007.

Lukács, G., & Ansorge, U. (2021). The mechanism of filler items in the response time concealed information test. *Psychological Research*, *85*(7), 2808–2828. https://doi.org/10.1007/s00426-020-01432-y.

Lukács, G., Kawai, C., Ansorge, U., & Fekete, A. (in press). Detecting concealed language knowledge via response times. *Applied Linguistics Review*. https://doi.org/10.1515/applirev-2020-0130.

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.

Lukács, G., & Specker, E. (2020). Dispersion matters: Diagnostics and control data computer simulation in Concealed Information Test studies. *PLOS ONE*, *15*(10). https://doi.org/10.1371/journal.pone.0240259. Article e0240259.

Meijer, E. H., Verschuere, B., Gamer, M., Merckelbach, H., & Ben-Shakhar, G. (2016). Deception detection with behavioral, autonomic, and neural measures: Conceptual and methodological considerations that warrant modesty. *Psychophysiology*, *53*(5), 593–604. https://doi.org/10.1111/psyp.12609.

Noordraven, E., & Verschuere, B. (2013). Predicting the sensitivity of the reaction time-based concealed information test. *Applied Cognitive Psychology*, *27*(3), 328–335. https://doi.org/10.1002/acp.2910.

Norman, D. G., Gunnell, D. A., Mrowiec, A. J., & Watson, D. G. (2020). Seen this scene? Scene recognition in the reaction-time concealed information test. *Memory & Cognition*, *48*(8), 1388–1402. https://doi.org/10.3758/s13421-020-01063-z.

Peirce, J., & MacAskill, M. (2018). *Building experiments in PsychoPy*. Sage.

Proctor, R. W., & Lu, C.-H. (1999). Processing irrelevant location information: Practice and transfer effects in choice-reaction tasks. *Memory & Cognition*, *27*(1), 63–77. https://doi.org/10.3758/BF03201214.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/.

Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC area, Cohen's d, and r. *Law and Human Behavior*, *29*(5), 615–620. https://doi.org/10.1007/s10979-005-6832-7.

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). https://doi.org/10.1186/1471-2105-12-77. Article 77.

Seymour, T. L., & Schumacher, E. H. (2009). Electromyographic evidence for response conflict in the exclude recognition task. *Cognitive, Affective, & Behavioral Neuroscience*, *9*(1), 71–82. https://doi.org/10.3758/CABN.9.1.71.

Sochat, V. (2018). The experiment factory: Reproducible experiment containers. *The Journal of Open Source Software*, *3*(22). https://doi.org/10.21105/joss.00521. Article 521.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643–662. https://doi.org/10.1037/h0054651.

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.

Tagliabue, M., Zorzi, M., Umiltà, C., & Bassignani, F. (2000). The role of long-term-memory and short-term-memory links in the Simon effect. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(2), 648–670. https://doi.org/10.1037/0096-1523.26.2.648.

Thompson, T. J., & Amy, B. (2009). The habituation effect in personnel security polygraph screening. *Polygraph*, *38*(3), 218–222.

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*. https://doi.org/10.3389/fpsyg.2012.00526. Article 526.

Varga, M., Visu-Petra, G., Miclea, M., & Buş, I. (2014). The RT-based concealed information test: An overview of current research and future perspectives. *Procedia - Social and Behavioral Sciences*, *127*, 681–685. https://doi.org/10.1016/j.sbspro.2014.03.335.

Vendemia, J., Buzan, R., & Green, E. (2005). Practice effects, workload, and reaction time in deception. *American Journal of Psychology*, *118*(3), 413–429.

Verschuere, B., & De Houwer, J. (2011). Detecting concealed information in less than a second: Response latency-based measures. In B. Verschuere, G. Ben-Shakhar and E. Meijer (Eds.), *Memory detection* (pp. 46–62). Cambridge University Press.

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.

Viechtbauer, W. (2010). Conducting meta-analyses in *R* with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03.

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. Article 146.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis* (2nd ed). Springer.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models: Estimation of Semiparametric Generalized Linear Models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(1), 3–36. https://doi.org/10.1111/j.1467-9868.2010.00749.x.

(*Appendices follow*)

Appendix

Optimal Trial Numbers

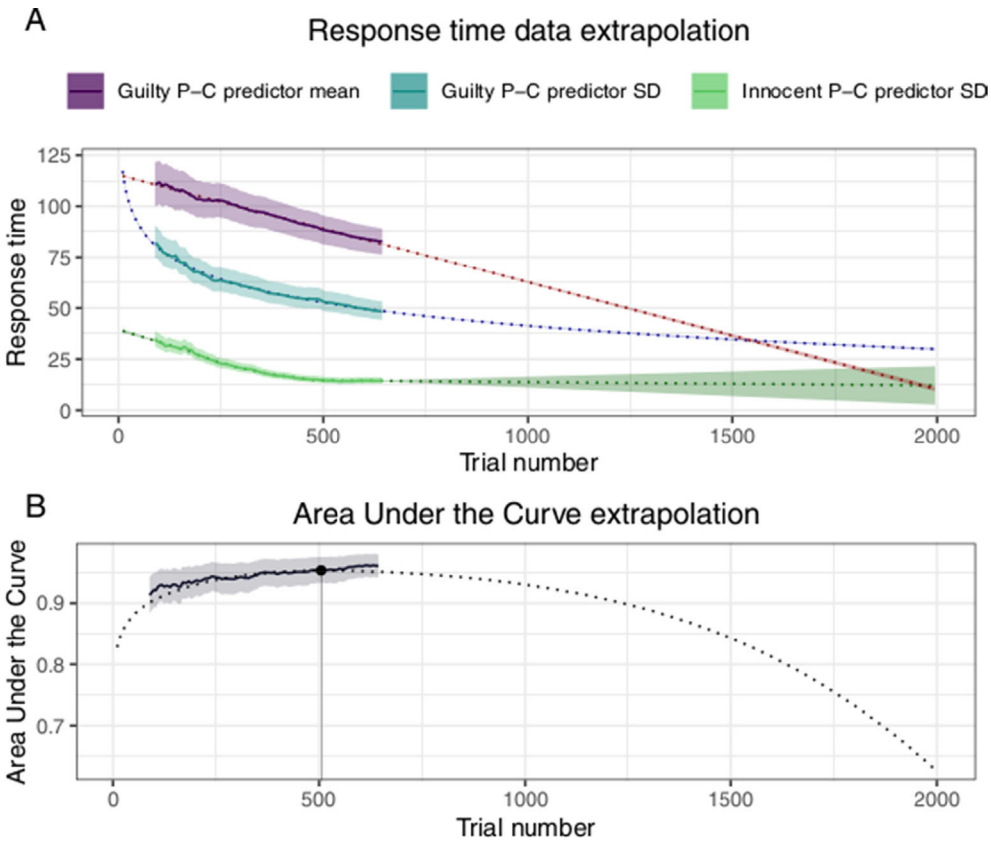
The essentially uninterrupted increase of the AUC values with the increasing number of trials, as seen in Figures 3 and 6, could be misleading. It would be surprising if AUC always increased up to perfect accuracy with longer testing. In particular, if the decrease of the group means of individual probe-control RT differences remains linear, it must at some point reach zero. From that point onward, the AUC can only decrease (as long as it is still above 0.5, i.e., chance level). Furthermore, a linear decrease of the SDs of probe-control RT difference toward zero seems unlikely (Lord & Novick, 2008), and most of the present data (Figure 2) also seems to indicate a nonlinear tendency.

Using combined data from Studies 2a and 2b shows that the linear regression for the probe-control difference means, and logistic regression for the probe-control difference SDs, both prove to be excellent fits; $R^2 = 0.92$, $F(1, 27) = 324.5$, $p < .001$, and $R^2 = 0.94$, $F(1, 27) = 454.9$, $p < .001$, respectively; see Figure A1, top Panel (A). The SDs of innocent probe-control differences can be similarly fitted (although, in this case, a generalized additive model was needed for a reasonable fit [Wood, 2011], which makes model-based predictions less reliable; Figure A1). Based on these models, predicted mean and SD can be obtained for any given trial number: Figure A1, top Panel (A), shows the predictions for the trials from zero to 2,000.

Finally, for any given trial number, a simulated guilty individual predictor variable (in lieu of empirical individual probe-control RT differences) can be generated as normally distributed data with the predicted means and SDs of guilty probe-control differences. Likewise, simulated innocent predictor variables can be generated as normally distributed data with the predicted SD of innocent probe-control difference and a mean of zero (assuming no probe

Figure A1

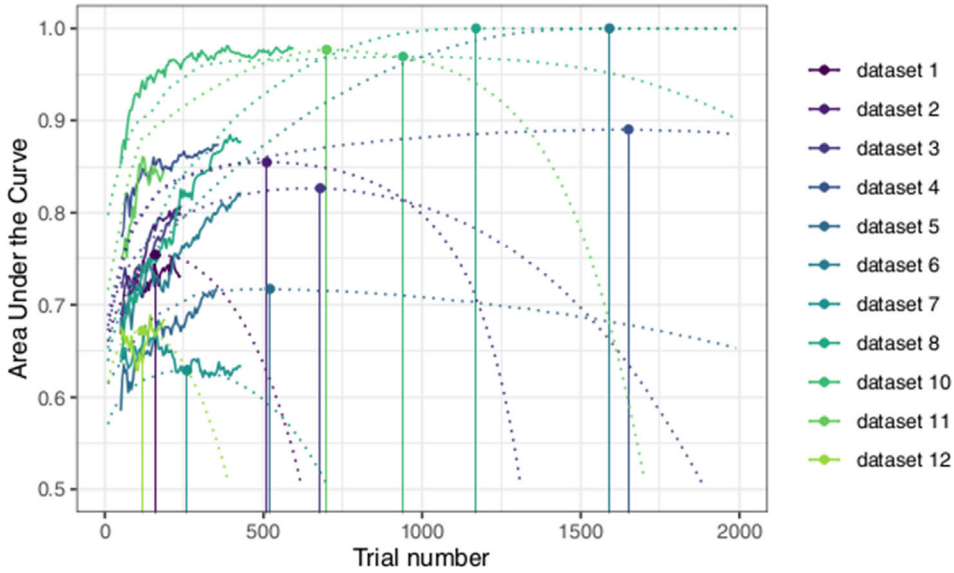
Model-Based Extrapolation of Cumulative RT and AUC Data



Note. Per trial number, the top Panel (A) depicts the means (guilty only) and SDs (guilty and innocent) of empirical probe-control (P-C) RT differences from real data (Study 2a and 2b merged; solid lines), as well as the computational extrapolation of the same values (dotted lines). The bottom Panel (B) correspondingly depicts the empirical AUCs from the real data (solid line), as well as simulated AUCs (dotted line) based on the extrapolated means and SDs. The filled circle (connected to the *x* axis with a horizontal line) marks the point where the maximum value of the simulated AUCs is reached. Ribbons around the lines depict 95% CIs. See the online article for the color version of this figure.

Figure A2

Model-Based Cumulative Areas Under the Curves



Note. Empirical AUCs from the real data (Study 2a and 2b merged; solid line), as well as simulated AUCs (dotted line) based on the extrapolated means and SDs of probe-control RT differences. Dataset 9 is omitted, because no reasonable fitting was possible on so few trials (see in Figure 2). The filled circles (connected to the *x* axis with horizontal lines) mark the points, per each given dataset, where the maximum value of the given dataset's simulated AUCs is reached. See the online article for the color version of this figure.

recognition, hence no probe-control RT difference on average). From the simulated guilty and innocent predictor variables (100 values per each) at each given trial, an AUC can be calculated (regarding the reliability of such simulated AUCs, see Lukács & Specker, 2020). Thereby the AUC may be indirectly extrapolated based on the underlying data; see Figure A1, bottom Panel (B). This extrapolation shows that, if the speculative bases of these procedures were accurate, the AUC would be expected to peak at around the trial range of 500–550, roughly coinciding with the maximum extent of the empirical data.

A similar procedure may be applied for any given new empirical RT-CIT data (from different task designs or experimental settings) to predict AUC values at an extended trial number. For illustration, such predictions were computed for the datasets from Study 1 and depicted in Figure A2. (Note that here in several datasets, the assumptions are even less certain due to the relatively small number of trials.)

The results seem to imply that the RT-CIT would reach its peak classification accuracy before 1,700 in all included experiments and that, thereby, the optimal trial numbers are generally within this range. However, the primary purpose of these model-based predictions is to provide a proof of principle for the possibility of eventually declining AUCs, and it is not certain to what extent these patterns may hold true in reality. For example, the assumption that probe-control RT differences will steadily continue to decline linearly all the way to zero seems unlikely. Therefore, the simulations above may be pessimistic regarding classification accuracy in case of prolonged testing. For a valid assessment, empirical data is needed using longer RT-CITs.

Received March 20, 2021
Revision received August 26, 2021
Accepted August 26, 2021 ■