UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Lukács, G., et al. (2020). Response Time Concealed Information Test on Smartphones. *Collabra: Psychology,* 6(1): 4. DOI: https://doi.org/10.1525/collabra.255

ORIGINAL RESEARCH REPORT

Response Time Concealed Information Test on Smartphones

Gáspár Lukács[*,†], Bennett Kleinberg[‡], Melissa Kunzi[*] and Ulrich Ansorge[*,§]

The Response Time-Based Concealed Information Test (RT-CIT) can reveal when a person recognizes a relevant (*probe*) item among other, irrelevant items, based on comparatively slower responding to the probe item. 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. So far, the RT-CIT has been used only on desktop computers. In Experiment 1 (n = 72; within-subject), we compare the probe-irrelevant differences when using the conventional desktop-based CIT to using a smartphone-based CIT, demonstrating practical equivalence. In Experiment 2 (n = 116; within-subject), we demonstrate that using thumbs for responses (while holding the smartphone) leads to equally efficient CIT results as using conventional index finger responses. At the same time, this second experiment also demonstrates how smartphone-based studies may be efficiently run in large groups, using the participants' own smartphones. Finally, as an interesting addition, here for the first time we also measured keypress durations (i.e., the time durations of holding down the response keys) in the RT-CIT, which we found to be significantly shorter for probe than for irrelevant items.

Keywords: deception; concealed information test; response time; smartphone; mobile

Undetected deception may lead to extreme costs in certain scenarios such as counterterrorism, pre-employment screening for intelligence agencies, or high-stakes criminal proceedings. However, meta-analyses have repeatedly shown that without special aid, based on their own best judgment only, people (including police officers, detectives, and professional judges) distinguish lies from truths on a level hardly better than mere chance (Bond & DePaulo, 2006; Hartwig & Bond, 2011; Kraut, 1980). Therefore, researchers have advocated special techniques that facilitate lie detection, among which the most prominent ones are information-elicitation interviewing techniques (e.g., Vrij & Granhag, 2012) and the use of technology (e.g., computerized tasks as in the present study).

One of the potential technological aids is the Concealed Information Test (CIT; Lykken, 1959; Meijer, Selle, Elber, & Ben-Shakhar, 2014). The CIT aims to disclose whether examinees recognize certain relevant items, such as a weapon used in a recent homicide, among a set of other objects, when they actually try to conceal any knowledge about the criminal case. In the response time (RT)-based CIT, participants classify the presented stimuli as the target or as one of several non-targets by pressing one of two keys (Seymour, Seifert, Shafto, & Mosmann, 2000; Suchotzki, Verschuere, Van Bockstaele, Ben-Shakhar, & Crombez, 2017; Varga, Visu-Petra, Miclea, & Buş, 2014). Typically, five non-targets are presented, among which one is the *probe*, which is an item that only a guilty person would recognize, and the rest are *irrelevants*, which are similar to the probe and, thus, indistinguishable from it for an innocent person. For example, in a murder case where the true murder weapon was a knife, the probe could be the word "knife," while irrelevants could be "gun," "rope," etc. Assuming that the innocent examinees are not informed about how the murder was committed, they would not know which of the items is the probe. The items are repeatedly shown in a random sequence, and all of them have to be responded to with the same response keys, except one arbitrary *target* – a randomly selected, originally also irrelevant item that has to be responded to with the other response key. Since guilty examinees recognize the probe as the relevant item in respect of the deception detection scenario, it will become unique among the irrelevants and in this respect more similar to the rarely occurring target (Lukács & Ansorge, 2019a). Due

* Department of Basic Psychological Research and Research Methods, University of Vienna, Vienna, AT

† Department of Psychology, Humboldt University of Berlin, Berlin, DE

‡ Department of Security and Crime Science, University College London, London, UK

§ Vienna Cognitive Science Hub, University of Vienna, Vienna, AT

Corresponding author: Gáspár Lukács (gaspar.lukacs@univie.ac.at)

to this conflict between instructed response classification of probes as non-targets on the one hand, and the probe's uniqueness and, thus, greater similarity to the alternative response classification as potential targets on the other hand, the response to the probe will be generally slower in comparison to the irrelevants (Seymour & Schumacher, 2009). Consequently, based on the probe-to-irrelevant RT differences, guilty (i.e., knowledgeable) examinees can be distinguished from innocent (i.e., naive) examinees.

The RT-CIT takes little time (5–10 mins), its administration requires no special expertise,[1] and its results can be analyzed instantaneously, in a standardized way.

RT-CIT Application for Smartphones

The primary aim of our study was to show that the RT-CIT can be used just as well on a smartphone as on a desktop computer: This would provide a cost-free portable CIT lie detector that could be useful in various scenarios. Nonetheless, our study also has more general implications. In particular, while some recent studies have shown that RT tests may be validly administered on a smartphone (Burke et al., 2017; Kay et al., 2013; Schatz, Ybarra, & Leitner, 2015), this is the first study to directly compare RT task results on a smartphone versus on a desktop.

The main practical difference between smartphone and desktop is arguably the use of touchscreen versus physical keyboard (Kay et al., 2013). The former requires holding the finger hovering over the response surface, and tapping it as response in the task: very shortly touching the screen and then lifting up the finger. In the case of a desktop keyboard, the finger lies on the response key, which needs to be pressed down with force, afterwards letting it raise back by itself (i.e., by ceasing to exert force). Other potential differences may be hypothesized as well, such as increased allocation of attention to the screen due to the joint location of the presented stimuli and the response buttons in case of the smartphone, or the differing user experience associated with the given devices (i.e., people are habituated to different applications and corresponding finger movement patterns on a smartphone compared to a desktop). However, we presumed that all such potential differences are nonessential in respect of a task where the dependent variables are always based on a within-subject comparison (probe RT mean vs. irrelevant RT mean). Therefore, in Experiment 1, we directly compared the RT-CIT on *Smartphone* versus on *Desktop*, expecting no difference in the RT-CIT effect (probe-to-irrelevant RT differences in case of participants simulating guilty suspects).

Experiment 1

Method

The experiment was preregistered at https://osf.io/9cgjn/ (Foster & Deardorff, 2017; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012).

Participants

The tests were conducted at a behavioral experiment laboratory of the University of Vienna, where 77 psychology students completed our experiment (to receive experiment participation credits for curriculum requirements). The experiment was run in a within-subject design: Each participant completed once the Smartphone version, once the Desktop version. The test was taken in groups of two: While one participant was first tested in the Smartphone condition, the other was first tested in the Desktop condition, after which they did the reverse. All participants were tested with their own personal first and last names as probes in the CIT task, simulating a guilty suspect trying to conceal the recognition of these two names (see, e.g., Lukács, Gula, Szegedi-Hallgató, & Csifcsák, 2017; Verschuere & Kleinberg, 2015).

The data from the intended first two participants were excluded immediately after the completion of the test, due to small technical issues. The preregistered number of 75 participants were collected subsequently. Out of these, two participants were excluded due to entering, to be used as a probe, a double first name (i.e., including a middle name; despite our warning). Our exclusion criteria were an accuracy rate not over 50% for targets or not over 75% for main items (probe or irrelevant items). There was only one related exclusion (due to too low target accuracy). This left 72 valid participants ($M_{age} \pm SD_{age} = 21.14 \pm 1.65$; 12 male; 36 started with Smartphone).

Procedure

Before the beginning of the experiment, each participant read and signed an informed consent form, which also included the information that the following task simulates a lie detection scenario, during which participants should try to hide their identities.

For the CIT that came next, both the smartphone and the desktop applications were written in HTML5/JavaScript (for the use of this framework in RT tasks in general, see Reimers & Stewart, 2015; or in RT-CIT in specific, see Kleinberg & Verschuere, 2015). The Desktop version was run in Google Chrome (Version 70.0.3538), while for the Smartphone version the same code was adapted into the Ionic Framework to create a hybrid mobile application (built for Android; see e.g., Khandeparkar, Gupta, & Sindhya, 2015). This latter allowed the implementation of several useful native smartphone functionalities; in particular true full screen (no interfering notification or navigation bars) and automatic local storage of the resulting data.

The consequent tests on the Smartphone and Desktop were identical, except for the following points. In case of the keyboard (Logitech K120 920-003626) of the desktop computer, a simple keypress was required: key "F" as response on the left, and key "K" as response on the right. (For corresponding response categories, see below.) In case of the touchscreen of the smartphone (Moto G5 XT1676), a *tap* was required as response: the touching of the screen (finger down) and releasing it (finger lifted up; within 300 ms of the touch start). The layout of the two response fields was designed to have approximately the same size (and form) as the surface of the keyboard keys. The distance between the left and right response fields was the same as the distance between the left and right keyboard keys (ca. 6.5 cm; surface size per field or key: ca. 2.4 × 2.2 cm). The keyboard and the smartphone were switched in the

same place after the first test, so that, in the second test, the position of the responses (keys or fields) remained the same. In case of the keyboard, the key letters ("F," "K") were mentioned only once in the beginning of the task; afterwards, same as in case of Smartphone, the responses were referred to only as *left-side* or *right-side* response. In either case, participants used their left and right index fingers to respond. In case of Desktop, the monitor was placed next to the keyboard, and the items in the CIT task appeared in the middle of its screen (37.5 × 30.0 cm). In case of the Smartphone screen (11.0 × 6.2 cm; always in horizontal mode), the items appeared above the response fields, in the top half of the screen (see **Figure 1**). Both screens were the same distance from the eyes of the participant (ca. 55 cm). Consequently, participants looked at the smartphone screen in a larger angle from horizontal (ca. 50°) than at the desktop screen (ca. 24°). Each detail of the rest of the description, as follows below, applies to both versions equally.

Participants entered their first names (along with gender) and last names, which then served as the two probe items in the task.[2] For each probe, five items were randomly chosen from a list of frequent German names (with corresponding gender for first names; 117 female and 138 male first names; 100 last names), out of which one was randomly chosen to serve as target, while the remaining four served as irrelevants. The random choice was restricted in that these five items had the closest possible character length to the given probe, and not any two of the six items started with the same letter. Thus, for each participant, there were altogether 12 unique items: two probes, two targets, and eight irrelevants (all 12 identical in the Desktop and Smartphone conditions.) We refer to the probes and irrelevants jointly as non-targets.

Next, participants were presented the two target names, and were asked to memorize these items in order 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 entered these items correctly. If any of the entered items was incorrect, the participant received a warning and was redirected to the previous page in order to have another look at the same items.

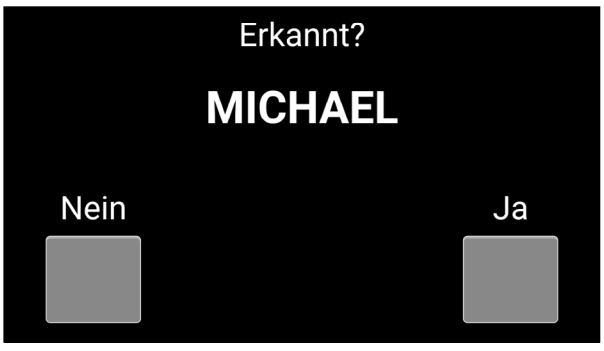

Figure 1: Example screenshot from the first practice phase of the Smartphone-based CIT, with "MICHAEL" displayed as the item awaiting a response. (In the following practice and main tasks, the reminder captions [*Erkannt?* – "Recognized?"; *Nein* – "No"; *Ja* – "Yes"] were not displayed anymore.)

During the task, the items were presented one by one on the screen (in 0.65 cm tall uppercase letters; in white font color on black background) and participants had to categorize them with one of the two response alternatives. Participants were told that the right-side response means "Yes," they recognize the item, while the left-side response means "No," they do not recognize the item – and they were correspondingly instructed to say "Yes" to the targets, and "No" to all other, non-target words (i.e., both the irrelevants and the probes).

We implemented the simplest version of the RT-based CIT (single-probe protocol; see Verschuere, Kleinberg, & Theocharidou, 2015). We decided for this version in order to focus exclusively on the most general aspect of the CIT, namely the recognition of a single relevant detail among irrelevant details, without any additional complexities that are involved in more efficient protocols (e.g., Lukács & Ansorge, 2019b; Lukács, Kleinberg, & Verschuere, 2017; Verschuere & Kleinberg, 2015).

More precisely, there are two obvious superior alternatives: the multiple-probe protocol (Verschuere et al., 2015) and the single-probe protocol enhanced with familiarity-related filler items (Lukács, Kleinberg, et al., 2017; or the even more complex related alternatives; Lukács & Ansorge, 2019b). The former mixes all item categories within the CIT (e.g., first and last names) randomly together in one block, while the latter includes additional familiarity- and unfamiliarity-referring items (e.g., "familiar" and "unknown"). Thus, in both cases the trial sequence randomization can result in an unequal distribution of preceding items for any given item (e.g., several targets preceding a probe), whose semantic priming effect may influence the response speed to the upcoming word (e.g., Foss, 1982; Meyer & Schvaneveldt, 1971), hence, introducing theoretically uninteresting statistical noise in the data. Furthermore, the single-probe protocol has several practical advantages as compared to the multiple-probe protocol (as detailed by Lukács, Kleinberg, et al., 2017): applicability even in case of a limited number of probe items (Podlesny, 2003), compatibility with common test procedures and scoring algorithms (Krapohl, 2011), and sequential testing to narrow down possibilities (Lukács, Kleinberg, et al., 2017), and that (consequently) practitioners currently also consider the single-probe protocol to be the only viable option (Ogawa, Matsuda, Tsuneoka, & Verschuere, 2015). The addition of familiarity-related fillers, on the other hand, is also a recent development, and we preferred to use a well-established CIT protocol.

In sum, we found it best to use the simplest possible protocol, which is the single-probe protocol. We presume that the difference between the use of smartphone versus desktop involves very basic cognitive and behavioral processes (mainly: quite simply the pressing of key vs. touching a touchscreen) that could hardly be affected by specific CIT versions – therefore, the outcome of the comparison can be subsequently extrapolated to any of the more complex designs. (Relatedly, to ensure that we still obtain large enough probe versus irrelevant effects to be compared between devices, we chose very high salient probes, namely, the participants' personal names.)

During the comprehension check and the first practice task (see below), reminder captions were displayed: "Recognized?" (*Erkannt?*) at the top of the screen, and, in the lower part of the screen, "No" (*Nein*) on the left and "Yes" (*Ja*) on the right (**Figure 1**). Starting from the second practice task (and throughout the main blocks), these captions were not displayed anymore.

The inter-trial interval (i.e., between the end of one trial and the beginning of the next) always randomly varied between 300 and 800 ms. In case of a correct response, the next trial followed. In case of an incorrect response or no response within the given time limit, the caption "Wrong!" (*Falsch!*) or "Too slow" (*Zu langsam!*) appeared, respectively, below the stimulus in red color 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. All 12 items were displayed in a random order, and participants had plenty of time (10.5 s) to choose a response – however, each trial required a correct response. In case of an incorrect response, the participant immediately got a corresponding feedback, was reminded of the instructions, and had to repeat this check. This check guaranteed that the eventual differences (if any) between the responses to the probe and the responses to the irrelevants were not due to misunderstanding of the instructions or any uncertainty about the required responses in the eventual task.

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 12 trials (first the six items from one name category, then the six from the other; in the order of the main blocks; see below). In either practice task, in case of too few valid responses, the participants received a corresponding feedback, were reminded of the instructions, and had to repeat the practice task. The requirement was a minimum of 60% valid responses (correct response between 150 and 800 ms) for targets and for main items (probes and irrelevants together).

The main task contained two blocks: one with first names only, and one with last names only (order counterbalanced across participants, but, for each participant, same order in each condition, Desktop and Smartphone). Each probe, irrelevant, and target was repeated 18 times in each block (hence, altogether 36 probe, 72 irrelevant, and 36 target trials). Within each block, 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).

After this test was completed in both conditions (Desktop and Smartphone), participants gave their demographic details and completed a very brief questionnaire regarding their alertness during the task (see Appendix A). Finally, participants were given more detailed information about the experiment and contact details for potential further inquiries. The experiment took about 30 min per session.

Data Analysis

We conducted preregistered analyses, except where explicitly noted otherwise.

For the main questions, the dependent variable was the probe-to-irrelevant RT mean (i.e., probe RT mean minus irrelevant RT mean, per each participant), which was compared between the Desktop and Smartphone conditions with three statistical tests: (a) a simple *t*-test to test the potential difference, (b) Bayesian likelihood ratio to test the null hypothesis, and (c) a *two one-sided t-test* (TOST) procedure, as a frequentist approach for testing the equivalence, with equivalence bounds of $d = -0.4$ and $d = 0.4$ (see below). Following the suggestion of a reviewer of a previous version of this manuscript, for probe-irrelevant RT means we report Spearman-Brown split-half reliability coefficients (Brown, 1910; Eisinga, Grotenhuis, & Pelzer, 2013; Spearman, 1910; for CIT, Kleinberg & Verschuere, 2015).

While in the RT-CIT usually only RT means are used as predictors (for guilty-innocent classifications), certain extents of probe-to-irrelevant differences are also often observed in accuracy rates as well, and therefore may be of interest (in particular, see Lukács & Ansorge, 2019b; Lukács, Gula, et al., 2017). Consequently, the three tests above were repeated with probe-to-irrelevant accuracy rate differences (i.e., probe accuracy rate minus irrelevant accuracy rate, per each participant), in place of RT means, as dependent variables.

In the preregistration, we only mentioned comparing keypress- and touch-durations (from here on, we designate these collectively as *hold-durations*), and the potential effects of self-reported alertness, as potential exploratory analyses. Here, we specify that, for hold-durations, we decided for an analysis of variance (ANOVA) with the two factors Trial Type (probe vs. irrelevant) and Device (Desktop vs. Smartphone).[3]

Regarding the alertness questionnaire, we tested the correlations of the aggregated ratings, in case of desktop and smartphone separately, with (a) probe-to-irrelevant RT mean differences, and (b) probe-to-irrelevant accuracy rate differences; these analyses are reported in Appendix A.

Finally, as a preliminary assessment of the potential incremental benefit of hold-durations, we report exploratory binary logistic regression analysis combining RT means and hold-durations, and present illustrative simulated areas under the curves (AUCs; see below) based on the fitted values. For each simulation, to represent the hypothetical "innocent" (or "naive") suspect's data, we generated a sample of 1,000 values in perfect normal distribution with a mean of zero (see in the R script uploaded to the OSF repository). In case of each predictor, we used the *SD* of the same predictor from the real participants' data.[4] For example, the probe-irrelevant RT mean *SD* in the Desktop condition was 27.1 ms; hence, the

simulated data was a normally distributed 1,000 values with *SD* = 27.1 (and a mean of zero).

Bayesian analysis

We report Bayes factors using the default *r*-scale of 0.707 (Morey & Rouder, 2018). The Bayes factor is a ratio between the likelihood of the data fitting under the null hypothesis and the likelihood of fitting under the alternative hypothesis (Jarosz & Wiley, 2014; Wagenmakers, 2007). For example, a Bayes factor (BF) of 3 means that the obtained data is three times as likely to be observed if the alternative hypothesis is true, while a BF of 0.5 means that the obtained data is twice as likely to be observed if the null hypothesis is true. Here, for more readily interpretable numbers, we denote Bayesian factors as BF_{10} for supporting alternative hypothesis, and as BF_{01} for supporting null hypothesis. Thus, for example, $BF_{01} = 2$ again means that the obtained data is twice as likely under the null hypotheses than under the alternative hypothesis. Typically, BF = 3 is interpreted as the minimum likelihood ratio for "substantial" evidence for either the null or the alternative hypothesis (Jeffreys, 1961).

TOST

In the TOST procedure, the null hypothesis, analogous to a simple *t*-test, is the *presence* of a true difference in either direction, with the effect sizes specified as the equivalence bounds, in our case $d = 0.4$ in either direction. If the *p* value for the one-sided *t*-tests examining either direction (or both) is below the alpha level (.05), we can assume that, in the given direction, there is no difference larger than the specified effect size (Lakens, 2017; Schuirmann, 1987).[5] As described in our preregistration, the conventional medium effect size of $d = 0.5$ has been shown, in previous studies, to be a reasonable practical indication of substantially increased CIT efficiency (e.g., Lukács & Ansorge, 2019b; Lukács, Kleinberg, et al., 2017; Verschuere et al., 2015). Therefore, for an insubstantial difference, we chose a somewhat lower effect size. To note, we aim to reveal whether there is an equivalence within these bounds of $d = -0.4$ and $d = 0.4$, but this is not to say that differences smaller than that are always fully negligible in all respects – however, this is a reasonable estimation for the potential usefulness of the smartphone-based alternative of the RT-CIT.

AUCs

To illustrate the potential efficiency of discriminating between guilty and innocent suspects, we calculated AUCs (a diagnostic efficiency measure, for binary classification, that takes into account the distribution of all predictor values; Rice & Harris, 2005; Zou, O'Malley, & Mauri, 2007) for receiver operating characteristics (ROCs). The AUC can range from 0 to 1, where .5 means chance level classification, and 1 means flawless classification (i.e., all guilty and informed innocent classifications can be correctly made based on the given predictor variable, at a given cutoff point).

Effect sizes

To demonstrate the magnitude of the observed effects, for *F*-tests we report generalized eta squared (η_G^2) and partial eta squared (η_p^2) with 90% CIs (Lakens, 2013). We report Welch-corrected *t*-tests (Delacre, Lakens, & Leys, 2017), with corresponding Cohen's *d* values as standardized mean differences and their 95% CIs (Lakens, 2013). In case of TOST, we also report 90% CIs to show the effect size bounds at alpha level. We used the conventional alpha level of .05 for all statistical significance tests.

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 an RT below 150 ms).

All analyses were conducted in R (R Core Team, 2019; via Kelley, 2019; Lawrence, 2016; Lukács, 2019; Makowski, Ben-Shachar, & Lüdecke, 2019; Morey & Rouder, 2018).

Results

Aggregated RT mean, accuracy rate, and hold-duration, for the different stimulus types in each condition (Desktop and Smartphone), are given in **Table 1**, along with related effect sizes.

RT means

The *t*-test between the probe-to-irrelevant RT means of Desktop and Smartphone indicated no significant difference, $t(71) = -0.18$, $p = .860$, $d_{within} = -0.02$, 95% CI [–0.25, 0.21], 90% CI [–0.21, 0.17]. Bayesian hypothesis testing indicated substantial evidence in favor of the null-hypothesis, $BF_{01} = 7.60$. The TOST showed that the 90% CI of the effect is well within the equivalence bounds ($d = -0.4$ and $d = 0.4$): The one-sided *t*-test against the upper bound (null hypothesis of larger probe-to-irrelevant RT means for Desktop than for Smartphone) was significant, $t(71) = -3.57$, $p < .001$, as well as the one against the lower bound (null hypothesis of larger values for Smartphone), $t(71) = 3.22$, $p < .001$. The reliability coefficients were $\rho = .549$ for Desktop, and $\rho = .695$ for Smartphone.

Accuracy rates

The *t*-test between the probe-to-irrelevant accuracy rates of Desktop and Smartphone indicated no significant difference, $t(71) = 0.77$, $p = .443$, $d_{within} = 0.09$, 95% CI [–0.14, 0.32], 90% CI [–0.10, 0.28], and $BF_{01} = 5.80$. The TOST again showed that the 90% CI of the effect is well within the equivalence bounds ($d = -0.4$ and $d = 0.4$): The one-sided *t*-test against the upper bound (null hypothesis of larger values for Desktop) was significant, $t(71) = -2.62$, $p = .005$, as well as the one against the lower bound (null hypothesis of larger values for Smartphone), $t(71) = 4.17$, $p < .001$.

Exploratory analysis: Hold-durations

The ANOVA for hold-durations as dependent variable, with within-subject factors Trial Type and Device, revealed significant main effects for Trial Type (shorter duration for

Table 1: RT Means, Accuracy Rates, and Hold-Durations, in Experiment 1.

	RT mean		Accuracy rate		Hold-duration	
	Desktop	**Smartphone**	**Desktop**	**Smartphone**	**Desktop**	**Smartphone**
Probe	441 ± 55	496 ± 45	98.0 ± 3.4	97.6 ± 3.6	113 ± 23	86 ± 22
Irrelevant	404 ± 44	458 ± 33	99.2 ± 0.8	99.2 ± 1.0	114 ± 24	87 ± 22
Target	499 ± 43	573 ± 33	83.1 ± 10.7	86.1 ± 9.4	112 ± 23	80 ± 19
P – I	37.2 ± 27.1	37.8 ± 27.6	–1.26 ± 3.46	–1.59 ± 3.41	–1.8 ± 5.1	–0.9 ± 3.9
d_{PvsI}	1.38 [1.05, 1.70]	1.37 [1.04, 1.69]	–0.36 [–0.60, –0.12]	–0.46 [–0.71, –0.22]	–0.35 [–0.58, –0.11]	–0.24 [–0.47, 0.00]
AUC	.834	.839	.579	.607	.596	.574

Note: Means and *SD*s (in the format of M ± *SD*) for individual RT means, accuracy rates, and hold-durations; for *Probe* (participants' own names), *Irrelevant* (other names), *Target* (the designated irrelevant details that require different response), *P* − *I* (individual probe minus irrelevant values); separately for the *Smartphone* and *Desktop* computer conditions. Cohen's *d* effect sizes (as d_{PvsI}) and simulated *AUC*s for the probe-to-irrelevant differences are given under each respective column.

probe items), $F(1, 71) = 10.39$, $p = .002$, $\eta_p^2 = .128$, 90% CI [.030, .249], $\eta_G^2 = .001$, as well as for Device (shorter duration for Smartphone), $F(1, 71) = 121.42$, $p < .001$, $\eta_p^2 = .631$, 90% CI [.512, .706], $\eta_G^2 = .263$. The significant Trial Type main effect may be surprising considering the small real difference (1–2 ms, see **Table 1**), but it is explained by the extremely high correlation of probe and irrelevant hold-durations: $r(70) = .977$, 95% CI [.964, .986] for Desktop, $r(70) = .983$, 95% CI [.974, .990] for Smartphone. (That is: the raw time difference may appear very small [as compared to, e.g., the RT mean differences], but, even relative to its millisecond magnitude, it is very consistent across participants.) We found no Trial Type × Device interaction, $F(1, 71) = 1.54$, $p = .218$, $\eta_p^2 = .021$, 90% CI [0, .103], $\eta_G^2 < .001$.

Exploratory analysis: Logistic model-based predictors

Using probe-irrelevant RT mean differences and probe-irrelevant hold-duration differences as two potential predictors in a logistic regression model, we fitted values in order to assess the potential incremental value of hold-durations in predicting the (simulated) conditions of guilt and innocence. The assessment of goodness-of-fit revealed a significant improvement relative to a constant-only model for both conditions, Desktop: $\chi^2(2) = 60.6$, $p < .001$, and Smartphone: $\chi^2(2) = 71.3$, $p < .001$. In both cases, the outcome (guilt vs. innocence) was significantly associated with both RT means (Desktop: $B = 0.25$, χ^2 [1] = 60.0, $p < .001$; Smartphone: $B = 0.16$, χ^2 [1] = 70.7, $p < .001$) and, importantly, hold-durations (Desktop: $B = -1.21$, χ^2 [1] = 52.2, $p < .001$; Smartphone: $B = -0.97$, χ^2 [1] = 54.3, $p < .001$), as individually contributing predictors, meaning that the inclusion of hold-duration did lead to significant improvement in predictions. Additionally, with a likelihood-ratio test we compared the model including only RT means with the model combining RT means and hold-durations: The latter model proved to be a statistically significant improvement in case of both versions (Desktop: χ^2 [1] = 222.7, $p < .001$; Smartphone: χ^2 [1] = 184.0, $p < .001$).

The AUC for the model-based predictors (fitted values) was .903 for Desktop, and .843 for Smartphone.

Discussion

Using a single-probe protocol RT-CIT with the participants' first and last names as probes, Experiment 1 has shown that, as hypothesized, the smartphone-based version can be as efficient as the desktop-based version. There is, however, an additional aspect of using smartphone as compared to desktop computer: namely, the touchscreen of the smartphone can also easily be operated with thumbs instead of index fingers; which allows more mobility for the user, and which is in fact the more common and natural way for smartphone usage in general (e.g., Azenkot & Zhai, 2012; Bröhl, Mertens, & Ziefle, 2017). Several studies have shown that using index finger responses instead of thumbs can lead to different results: In particular, it has been consistently found that more accurate general input (mainly: typing) can be given using index fingers (Buschek, De Luca, & Alt, 2016; Lehmann & Kipp, 2018; Wang & Ren, 2009; Wobbrock, Myers, & Aung, 2008). However, results have been mixed regarding speed differences, which seems to depend on the particular input type and study design (Azenkot & Zhai, 2012; Goel, Jansen, Mandel, Patel, & Wobbrock, 2013; Lehmann & Kipp, 2018; Wobbrock et al., 2008). In any case, to our knowledge, no studies have explored the potential effect of this difference in a regular experimental RT task yet, let alone in the RT-CIT. Therefore, in Experiment 2, we compared the conditions of using index fingers (*Index finger* condition) and of using thumbs (*Thumb* condition), to see whether the latter is at least as efficient as the former (i.e., yielding at least as large probe-to-irrelevant differences).

In Experiment 1, we also had a novel, exploratory finding: shorter hold-durations for probe compared to irrelevant items. This finding was significant in both conditions – though with a rather small effect, especially in the smartphone conditions. Therefore, an additional reason for Experiment 2 was to replicate this novel finding; with a larger sample size, and now, based on Experiment 1, with a data-based prediction before the experiment.

Finally, in Experiment 2, we also demonstrate how smartphone-based experiments may be efficiently run in larger groups, using the participants' own smartphones.

Experiment 2

Method

The experiment was preregistered at https://osf.io/g98nm/.

Participants

Participants were recruited similarly as in Experiment 1 (but those from Experiment 1 were not allowed to participate again). The difference was that participants brought their own smartphones for performing the test. Consequently, larger groups could be tested at once. For the first test, only six participants were invited (five signed up and were tested in a laboratory) as a final assurance of the technical feasibility before inviting larger groups. Since there were no issues, these participants were also included in the final sample. For all following tests, participants were invited in groups of 20, and the tests were conducted in a small classroom. (Average turnout was 11.1 persons per session; somewhat lower than expected, presumably due to the concurrent exam period.)

To avoid unexpected issues and facilitate the experiment procedure, participants were asked to (1) download, install, and make a quick pretest[6] with a small application similar to the one used in the experiment, and (2) download the application for the experiment in advance. In those relatively few cases when a participant did not make these preparations, they were still allowed to participate with their own smartphone if it had an Android operating system (OS)[7] – or otherwise they were provided with one (14 out of the 116 cases).

Again, the experiment was run with a within-subject design: Each participant completed the test once using their index finder, once using their thumbs (with order counterbalanced across participants). Again, all participants were tested with their own personal first and last names as probes, simulating a guilty suspect.

We stopped opening new slots for participants when the participant number first passed the preregistered number of 112. Up to that point, 116 participants completed the task ($M_{age} \pm SD_{age} = 21.15 \pm 2.42$; 43 male; 58 started with Index finger), none of whom had to be excluded.[8]

Procedure

Same as in Experiment 1, before the beginning of the experiment, each participant read and signed an informed consent form. Here, they also received an additional sheet with short instructions on how to download the application in case they had not already done that. Participants were also warned on this instruction sheet to disable all data connections and switch on airplane mode before starting the test.[9]

Next, the screen size of each smartphone was measured and entered on the start screen of the application – for all related details, see Appendix B. (The subject number was also entered at the same time: These numbers were printed on the information sheet, and the order of conditions and blocks were afterwards automatically assigned in the application based on the entered numbers: start with Index finger for every second participant; start with last names for every third and fourth). When every participant present at the session reached that point and was ready to start, the experimenter went around to make the application, using a hidden swipe area, move on to the rest of the test (the actual CIT).

The CIT was exactly the same as described in Experiment 1, except for the two conditions (Index finger vs. Thumb instead of Desktop vs. Smartphone), and some minor details, as follows. The sizes of the displays, including all shapes and text (captions, stimuli) were automatically sized relative to the smartphone on which the application was opened. (Note though, that this size variance is practically imperceptible in the CIT: Even for a difference between a 4.5- and a 5.5-inch screen, the stimulus height varies only as 0.59 and 0.72 cm, i.e., hardly more than a one-millimeter difference.)

After the completion of the first test (one block with first names, one with last names, in random order, in either Index finger or Thumb condition), each participant was asked to change their hand position, while keeping the smartphone approximately in the same position. In the Index finger condition, the smartphone lay on the desk, with participants using their index fingers for responding. In the Thumb condition, participants used their thumbs for responding, holding the smartphone in their hands, while their hand lay on the desk. Unlike Experiment 1, only the last practice phase was repeated after changing conditions (from Index finger to Thumb, or vice versa.) The use of correct condition was verified by an onlooking experimenter. (Participants were given consecutive subject numbers, and, thereby, consecutive orders of condition, by desk. Hence, someone using a wrong hand position would have been easy to notice.)

After the completion of the task in both conditions (Index finger and Thumb), each participant sent their results data file via email, using a button in the application that automatically prepared the message and included the data file as attachment.

On the end page, along with general information about the task, participants were also informed about whether they were classified as "guilty" or not (of concealing the recognition of their true personal names), based on their RTs. Here, for the first time, we used a proper automatic calculation for individual probe-to-irrelevant effect sizes, as $d_{CIT} = (M_{RT\ (probes)} - M_{RT\ (irrelevants)})/SD_{RT\ (irrelevants)}$ (from all four main blocks, valid trials only: correct responses with RT between 150 and 800 ms), with an arbitrary limit of minimum $d_{CIT} = 0.1$ to evaluate an outcome indicating concealed knowledge (Noordraven & Verschuere, 2013).

Participants were asked to remain silent at their place until everyone finished. The entire experiment, including all preparations, took about 30 min per session.

Data Analysis

We again conducted preregistered analyses (https://osf.io/g98nm/), except where explicitly noted otherwise.

For the main hypothesis (Thumb at least as efficient as Index finger), the dependent variable was the probe-to-irrelevant RT mean, which was compared between the Index finger and Thumb conditions with three statistical tests, analogously to Experiment 1: (1) a simple *t*-test,

(2) Bayesian hypothesis tests, and (3) a TOST procedure with equivalence bounds of $d = -0.4$ and $d = 0.4$. Again, these three tests were repeated with probe-to-irrelevant accuracy rates as dependent variables.

We preregistered the testing of hold-duration between probe and irrelevant items using a one-sided *t*-test, expecting shorter durations for probes (based on Experiment 1), along with a complementary Bayesian analysis. Here, we further specify that we do this separately for Index finger and Thumb conditions (and designate it as exploratory analysis). To provide more justification for this, we first perform an ANOVA, similarly as in Experiment 1, with the two factors Trial Type (probe vs. irrelevant) and Hand-position (Index finger vs. Thumb), to show whether there is an interaction.

Regarding correlation tests of the physical screen size with probe-to-irrelevant RT mean differences and with probe-to-irrelevant accuracy rate differences, see Appendix B.

Finally, as in Experiment 1, we report exploratory logistic regression analysis combining RT means and hold-durations, and present illustrative simulated AUCs.

Results

Aggregated RT mean, accuracy rate, hold-duration, for the different stimulus types in each condition (Index finger and Thumb), are given in **Table 2**, along with related effect sizes.

RT means

The *t*-test between the probe-to-irrelevant RT means of Index finger and Thumb conditions indicated no significant difference, with Bayesian hypothesis testing supporting the null hypothesis; $t(115) = 0.19$, $p = .848$, $d_{within} = 0.02$, 95% CI [–0.16, 0.20], 90% CI [–0.13, 0.17], $BF_{01} = 9.53$. The TOST showed that the 90% CI of the effect is well within the equivalence bounds ($d = -0.4$ and $d = 0.4$): The one-sided *t*-test against the upper bound (null hypothesis of larger probe-to-irrelevant RT means for Index finger than for Thumb) was significant, $t(115) = -3.57$, $p < .001$, as well as the one against the lower bound (null hypothesis of larger values for Thumb), $t(115) = 3.22$, $p < .001$. The reliability coefficients were $\rho = .739$ for Index finger, and $\rho = .608$ for Thumb.

Accuracy rates

Unlike in case of RT means, the *t*-test between the probe-to-irrelevant accuracy rates of Index finger and Thumb indicated a small but statistically significant difference, though with an inconclusive BF; $t(115) = 2.02$, $p = .046$, $d_{within} = 0.19$, 95% CI [0.00, 0.37], 90% CI [0.03, 0.34], $BF_{01} = 1.37$. The TOST again showed that the 90% CI of the effect is within the equivalence bounds ($d = -0.4$ and $d = 0.4$): The one-sided *t*-test against the upper bound (null hypothesis of larger values for Index finger) was significant, $t(115) = -2.29$, $p = .012$, as well as the one against the lower bound (null hypothesis of larger values for Thumb), $t(115) = 6.33$, $p < .001$. This altogether means that the accuracy rate difference between probe and irrelevant was statistically shown to be significantly larger in case of the Thumb condition, but at the same time, based on our predefined equivalence bounds, this difference is not of notable practical relevance.

Exploratory analysis: Hold-durations

The ANOVA for hold-durations as dependent variable, with within-subject factors Trial Type and Hand-position, revealed significant main effects for Trial Type (replicating shorter duration for probe items), $F(1, 115) = 4.66$, $p = .033$, $\eta_p^2 = .039$, 90% CI [.002, .111], $\eta_G^2 < .001$, as well as for Hand-position (shorter duration for Index finger), $F(1, 115) = 38.65$, $p < .001$, $\eta_p^2 = .252$, 90% CI [.143, .353], $\eta_G^2 = .066$. Here, we also found a significant Trial Type × Hand-position interaction, $F(1, 115) = 8.61$, $p = .004$, $\eta_p^2 = .070$, 90% CI [.013, .153], $\eta_G^2 < .001$, indicating larger probe-to-irrelevant differences for the Index finger condition. The follow-up one-sided *t*-tests indicated that the probe-to-irrelevant difference was significant only for Index finger, with BF strongly supporting this alternative hypothesis, $t(115) = -3.37$, $p < .001$, $d_{within} = -0.31$, 95% CI [–0.50, –0.13], $BF_{10} = 20.22$, and not for Thumb, with a

Table 2: RT Means, Accuracy Rates, and Hold-Durations, in Experiment 2.

	RT mean		Accuracy rate		Hold-duration	
	Index	**Thumb**	**Index**	**Thumb**	**Index**	**Thumb**
Probe	487 ± 54	478 ± 48	97.8 ± 3.3	97.3 ± 3.0	81 ± 21	93 ± 22
Irrelevant	451 ± 43	443 ± 39	98.8 ± 1.5	99.1 ± 1.2	82 ± 21	93 ± 22
Target	570 ± 53	550 ± 46	81.2 ± 10.9	80.2 ± 10.0	75 ± 19	94 ± 22
P – I	35.3 ± 29.0	34.9 ± 26.3	–0.99 ± 3.26	–1.76 ± 3.20	–1.4 ± 4.4	0.1 ± 3.9
d_{PvsI}	1.22 [0.97, 1.46]	1.33 [1.08, 1.58]	–0.30 [–0.49, –0.12]	–0.55 [–0.74, –0.35]	–0.31 [–0.50, –0.13]	0.03 [–0.16, 0.21]
AUC	.803	.828	.567	.640	.581	.499

Note: Means and *SD*s (in the format of M ± *SD*) for individual RT means, accuracy rates, and hold-durations; for *Probe* (participants' own names), *Irrelevant* (other names), *Target* (the designated irrelevant details that require different response), *P – I* (individual probe minus irrelevant values); separately for the Thumb (using thumbs) and Index (using index fingers) conditions. Cohen's *d* effect sizes (as d_{PvsI}) and simulated AUCs for the probe-to-irrelevant differences are given under each respective column.

BF supporting the null hypothesis, $t(115) = 0.29$, $p = .614$, $d_{within} = 0.03$, 95% CI [–0.16, 0.21], $BF_{01} = 9.31$.

Exploratory analysis: Logistic model-based classification

Using probe-irrelevant RT mean differences and probe-irrelevant hold-duration differences in a logistic regression model, we fitted values in order to assess the potential incremental value of hold-durations in predicting guilt or innocence. The assessment of goodness-of-fit revealed a significant improvement relative to a constant-only model for both conditions, Index finger: $\chi^2(2) = 90.9$, $p < .001$, and Thumb: $\chi^2(2) = 92.6$, $p < .001$. In both cases, the outcome was significantly associated with both RT means (Index finger: $B = 0.17$, χ^2 [1] = 90.1, $p < .001$; Thumb: $B = 0.15$, χ^2 [1] = 86.9, $p < .001$) and hold-durations (Index finger: $B = -1.05$, χ^2 [1] = 75.9, $p < .001$; Thumb: $B = -0.80$, χ^2 [1] = 60.8, $p < .001$), as individually contributing predictors. Additionally, with a likelihood-ratio test we compared the model including only RT means with the model combining RT means and hold-durations: The latter model proved to be a statistically significant improvement in case of both versions (Desktop: χ^2 [1] = 266.9, $p < .001$; Smartphone: χ^2 [1] = 169.2, $p < .001$).

The AUC for the model-based predictors was .857 for Index finger, and .838 for Thumb.

Discussion

In this second experiment we have shown that the hand position (using index fingers vs. thumbs for responses) plays no role in the results of the RT-CIT, at least regarding RT means. Regarding accuracy rates, we have shown that there is a small difference, in that slightly higher probe-to-irrelevant accuracy rate differences are found when using thumbs. This may be because using thumbs, as opposed to index fingers, is more sensitive to tasks requiring accuracy, and more prone to error rates in general (Buschek et al., 2016; Lehmann & Kipp, 2018; Wang & Ren, 2009; Wobbrock et al., 2008). This aspect could be explored in the future.

However, probe-to-irrelevant accuracy rate differences are in any case generally low in the RT-CIT, and have only rarely been used as predictors of guilt, but even in those cases only as secondary predictors. Nonetheless, if this aspect may be of any interest in the future, the method can still very well be always used with thumbs, as there is no general opposing reason or practical limitation. Note also that we have proven the equivalence, for the same accuracy rate differences, between the desktop and smartphone using index fingers. Consequently, the larger differences when using thumbs can only be an improvement (in respect of guilty-innocent predictions) as compared to the regular desktop version.

Our previous finding of shorter hold-durations for index finger was successfully replicated in this second experiment ($p < .001$). At the same time, this difference was absent in case of using thumbs. We have strong statistical support for this finding through both the ANOVA interaction (larger probe-to-irrelevant differences for Index finger; $p = .004$) and the Bayesian likelihood supporting the null finding (probe-to-irrelevant differences in case of Thumb; $BF_{01} = 9.31$). We also see a reasonable explanation for this. People are much more used to tapping with thumbs, as required by smartphone applications (for which usually thumbs are used): Touchscreens typically have a specific required hold-duration, only at the end of which is the given function executed (e.g., opening a folder). Hence, participants may be more strongly adapted to thumb *taps*, which are thereby more resistant to minor influences such as the probe-to-irrelevant differences in the RT-CIT. Nonetheless, this finding was not explicitly expected prior to the study and, therefore, would deserve further research.

General Discussion

In the present study, we have shown that the Response Time-Based Concealed Information Test (RT-CIT) can be used just as well on a smartphone as on a desktop computer. Before real life use, replication studies would be advisable, in particular field settings, and also including more efficient CIT protocols (Lukács, Kleinberg, et al., 2017; Verschuere et al., 2015). However, it already appears to be a valid method for various potential applications, facilitating the use of CIT in any situation where desktop computers are not available, limited, or impractical: such as border control (e.g., mass screening for the detection of country of origin[10]), pre-employment screening via remote interviews (where the smartphone application could automatically verify the device ID or phone number), or an immediately available test for appropriate investigating authorities, such as those in the police force, or in the military, at battlefronts (cf. the "handheld polygraph" of the U. S. Army; Dedman, 2008; Gordon, 2017; National Research Council, 2010; United States Office of the Secretary of Defense, 2018).

While not directly related to the main question of our study, we included an exploratory analysis in our first experiment on keypress- and touch-durations as topics relevant to other smartphone-based studies as well (e.g., Buschek, De Luca, & Alt, 2015; Goel et al., 2013). We found shorter durations for probes (i.e., when participants saw their own names), and replicated this finding in the second experiment (though only when using index fingers for touchscreen taps, and not when using thumbs). As compared to the use of RT mean alone, the combination of RT mean with hold-duration as model-based predictor led to noteworthy increases in classification efficiencies (AUCs) in two out of the four cases.

One reason for the duration differences could be that the lifting of the fingers corresponded to a second response and that some of the delay in the probe conditions was used to plan this second response in a sequence of responses consisting of key press and release (see Verwey, 1995). As a post-hoc test for this hypothesis, we calculated the correlations of response times and hold-durations per individual: These correlations were on average very weak (all correlation means between –.08 and .02) for both probe and irrelevant trial types in both conditions in both experiments – making the proposed hypothesis unlikely. Another potential explanation, however, is that participants perhaps felt their delay in the probe trials in

general (see Corallo, Sackur, Dehaene, & Sigman, 2008) and made an effort to compensate for the delay by a swifter key release. It could be interesting to explore whether this phenomenon appears in other RT tasks that contain target items, or any sort of response conflict or interference (e.g., Stroop tasks, cuing tasks, etc.).

Finally, while the feasibility of RT tasks on smartphones has been suggested before (Burke et al., 2017; Kay et al., 2013; Schatz et al., 2015), we provide strong evidence that such results can be identical to the ones obtained on regular computers. As demonstrated in our second experiment, using the participants' own smartphones, data can be easily collected in groups of 10–20, requiring nothing but an empty classroom. In the future, entire studies, with, say, over a hundred participants gathered in an auditorium, could be conducted this way within half an hour, with no equipment needed by the researchers. This would be a great advantage especially for less wealthy, less well-equipped universities and research institutes anywhere in the world.

Limitations

Our probe versus irrelevant effect sizes and simulated classification rates probably do not reflect well those that would be obtained in real life cases. While the personal relevance of the presented self-related autobiographical details arguably also resembles the relevance of real-life incriminating items, the extent of applicability is yet to be explored. In a specific situation very similar to the one simulated in the present study, authorities may test the true identity of the person, in which case the results may be assumed comparable to those in our study (regarding higher stakes at hand, see Kleinberg & Verschuere, 2016). This is, however, likely not a frequent case. The relevance of the more probable crime-related items (such as a murder weapon), which may be contributed to by the various emotions related to the actually committed crime (guilt, suspense, etc.), would be very difficult to simulate in a controlled experiment, and may require field studies in the future. In general, more realistic settings would be needed for a proper assessment of classification efficiency, as opposed to the highly controlled laboratory studies such as the present one, and indeed as it is in most RT-CIT studies.

Importantly, the primary aim of this study was to assess whether the smartphone-based CIT could be as efficient as the desktop-based one, and this comparison does not depend on precise or realistic demonstration of classification efficiencies. There is, however, one finding that could be substantially influenced by these biases: Namely, the incremental contribution of the novel hold-duration measure. This measure is, as we explicitly stated, an exploratory finding whose efficiency, usefulness, and mechanism should be assessed in future studies.

Conclusions

In the present study, using a single-probe protocol RT-CIT with the participants' first and last names as probes, we have (a) demonstrated that the smartphone-based version can be just as well used as the desktop-based version – using, for responses, either index fingers or thumbs (thus, simply holding the device in the hand), (b) shown that responses to probes compared to irrelevants in the RT-CIT have shorter keypress- and touch-durations – a difference that may be used as additional predictor of concealed knowledge, and which may be explored in other psychological tests as well, and (c) demonstrated a large-group experimental procedure using participants' smartphones, which may be adopted for any computerized tasks for fast and costless data collection in future studies.

Data Accessibility Statement

The source codes for all three experimental tasks, along with all behavioral data (original as well as aggregated per participant) and the R scripts for the analyses, are available via https://osf.io/fjvna/. For the smartphone applications, the original executable files are also available.

Appendix A

Alertness questionnaire

In Experiment 1, the short questionnaire at the end of the task consisted of the following four question: (1) At the moment I feel alert. (*Im Moment fühle ich mich aufgeweckt.*), (2) I was very focused on the task. (*Ich war sehr auf die Aufgabe fokussiert.*), (3) I felt very awake (alert) before the test. (*Vor dem Test fühlte ich mich sehr wach.*), (4) It was easy for me to stay concentrated during the test. (*Es fiel mir leicht, während des Tests konzentriert zu bleiben.*)

Each question could be rated on a six-point scale; from "I absolutely disagree" (*ich stimme absolut nicht zu*) to "I absolutely agree" (*ich stimme absolut zu*). For the analysis, the answers were assigned the value from 1 to 6, correspondingly, and were taken as one average from the four questions by each participant.

Calculating with CIT results from the desktop condition, the correlation of the ratings ($M_{rating} \pm SD_{rating}$ = 4.08 ± 0.68) with the probe-to-irrelevant RT mean differences was significant, $r(70) = -.248$, 95% CI [–.454, –.018], $p = .035$; but not with the probe-to-irrelevant accuracy rate differences, $r(70) = -.082$, 95% CI [–.308, .015], $p = .492$. Regarding the smartphone condition, the correlation was neither significant with probe-to-irrelevant RT mean differences, $r(70) = -.096$, 95% CI [–.320, .139], $p = .423$; nor with the probe-to-irrelevant accuracy rate differences, $r(70) = .164$, 95% CI [–.071, .381], $p = .169$.

The negative correlation of self-reported alertness with probe-to-irrelevant RT mean differences may be logical: For less alert participants, the CIT poses a larger cognitive load, that is, they might find it more difficult to make a quick categorization of the probe in spite of the response conflict (Visu-Petra, Varga, Miclea, & Visu-Petra, 2013). Nonetheless, it is not clear why this was significant only in case of the desktop CIT results. Furthermore, the difference is weak and hardly below the alpha level; thus, it may be just accidental. This topic should be addressed more thoroughly with a dedicated study, using a more proper, standardized questionnaire, and possibly an experimental manipulation (e.g., having the test performed in the morning vs. in the late evening).

Appendix B

Screen size measure

In Experiment 2, since participants used their own phones, we expected the screen sizes of the devices to vary to some small extent. This could not affect the outcome of our main results, since our study had a within-subject design. However, it still seemed worthwhile to record this data for potential exploratory analyses, in particular to see whether it has any effect on the CIT task's probe-to-irrelevant differences (as we preregistered it as a last secondary analysis). Therefore, the handout instruction sheet also included a printed but real size ruler with a corresponding grid (see via https://osf.io/fjvna/), which participants could use to measure the side lengths of their smartphone screens – by laying the paper's grid area over the screen, in which the screen margins (sides) were highlighted in red (via the CIT application), and could therefore be clearly seen through the paper. The entered numbers were double-checked by the experimenter before moving on to the CIT task.

Nonetheless, this was of very minor interest, and therefore it was noticed only after having collected data from 75 participants that the screen size information was not written out to the data file, and, hence, was lost for all these participants. This mistake was then immediately corrected, and, thus, these values were correctly saved for the remaining 41 participants. The related results below are reported with this partial data.

The correlation of the screen sizes ($M_{size} \pm SD_{size}$ = 76.43 ± 9.89 cm^2) with the probe-to-irrelevant RT mean differences or accuracy rates were not significant in either condition: for RT means, with Index finger, $r(39) = .237$, 95% CI [–.126, .469], $p = .237$, with Thumb, $r(39) = .159$, 95% CI [–.156, .445], $p = .320$; for accuracy rates, with Index finger, $r(39) = -.061$, 95% CI [–.362, .252], $p = .706$, with Thumb, $r(39) = -.098$, 95% CI [–.393, .217], $p = .544$.

Our sample size and the corresponding statistical power are much less for this analysis than as originally calculated – however, since this comparison was not at all the main subject of our study, we simply report it here as a tentative supplementary information. In any case, while it cannot be ruled out that the device size affects results (see, e.g., Lakens, Schneider, Jostmann, & Schubert, 2011; Lin, 2013), this effect, as also supported to some extent by our present data, is very unlikely to be substantial for the relatively small variance between typical smartphone sizes. Nonetheless, this may also be addressed in the future with more direct manipulation: For example, a within-subject design, where each participant performs the CIT using a large, a medium size, and a small device.

Notes

[1] Of course, using the CIT does require the proper understanding of the method's rationale, in order to appropriately select test items and correctly interpret the results – which is, however, quite straightforward.

[2] We confirmed in a post-hoc analysis that, in the final dataset, the entered names were always identical in the Desktop and Smartphone tests.

[3] In case of RT mean and accuracy rates, we did not include Trial Type as a factor, but, merely for simplicity and clarity, we used (as preregistered) one probe-to-irrelevant difference value instead of including probe and irrelevant results separately. (Note that separate inclusion of probe and irrelevant values gives identical results.) This was done because, unlike for hold-durations, there is ample unanimous evidence regarding the significant RT mean and accuracy rate differences between probes and irrelevants.

[4] Variations of this procedure make very little difference in the obtained AUCs.

[5] From another perspective, we can also say: We assume equivalence when the 90% CI of the effect size is within the specified bounds. The lower and upper limits of the 90% CI are therefore the points where the TOST test yields $p = .05$ for the given comparison. Hence, with $\alpha = .05$, using any bound within these intervals will be statistically nonsignificant ($p > .05$), while using any bound outside these intervals will be statistically significant ($p < .05$).

[6] The pretest consisted merely of opening the application and then creating an email message via a button click inside the opened application.

[7] To simplify both the development testing and the experiment procedure, we deployed the application for Android OS only. However, via the Ionic Framework, the same source code can just as well be deployed for iOS, or even for Microsoft Windows.

[8] Same as in Experiment 1, our exclusion criteria were an accuracy rate not over 50% for targets or not over 75% for main items, but no one in Experiment 2 violated these rules.

[9] If, on opening the application, internet connection was nonetheless detected, the participants were warned again automatically by an alert prompt within the application.

[10] The use of the CIT in this context would be most straightforward when there is only one, or at least only few, suspected place(s) of origin (see various examples by, e.g., McNamara, Van Den Hazelkamp, & Verrips, 2016), which could then be used as probes. It is, however, theoretically also possible to simultaneously test for a larger number of items in an "unknown probe scenario," with all test items as potential probes (for details, see, e.g., Meixner & Rosenfeld, 2011), although the diagnostic accuracy of such a test is yet to be demonstrated.

Funding Information

Gáspár Lukács is a recipient of a DOC Fellowship of the Austrian Academy of Sciences at the Institute for Basic Psychological Research and Research Methods at the University of Vienna.

Competing Interests

The authors have no competing interests to declare.

Author Contributions

Concept by B. K. and G. L.; study design and software by G. L.; acquisition of data by M. K. and G. L.; statistical analyses by G. L. and B. K.; manuscript drafted by G. L., revised by

U. A. and B. K, proofread by M. K. All authors approved the submitted version for publication.

References

Azenkot, S., & **Zhai, S.** (2012). Touch behavior with different postures on soft smartphone keyboards. In *Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services – Mobile HCI '12* (p. 251). San Francisco, California, USA: ACM Press. DOI: https://doi.org/10.1145/2371574.2371612

Bond, C. F., & **DePaulo, B. M.** (2006). Accuracy of deception judgments. *Personality and Social Psychology Review*, *10*(3), 214–234. DOI: https://doi.org/10.1207/s15327957pspr1003_2

Bröhl, C., Mertens, A., & **Ziefle, M.** (2017). How do users interact with mobile devices? An analysis of handheld positions for different technology generations. In J. Zhou & G. Salvendy (Eds.), *Human aspects of IT for the aged population. Applications, services and contexts* (Vol. 10298, pp. 3–16). Cham, Germany: Springer International Publishing. DOI: https://doi.org/10.1007/978-3-319-58536-9_1

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 1904–1920, 3*(3), 296–322. DOI: https://doi.org/10.1111/j.2044-8295.1910.tb00207.x

Burke, D., Linder, S., Hirsch, J., Dey, T., Kana, D., Ringenbach, S., ... Alberts, J. (2017). Characterizing information processing with a mobile device: Measurement of simple and choice reaction time. *Assessment*, *24*(7), 885–895. DOI: https://doi.org/10.1177/1073191116633752

Buschek, D., De Luca, A., & **Alt, F.** (2015). Improving accuracy, applicability and usability of keystroke biometrics on mobile touchscreen devices. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems – CHI '15* (pp. 1393–1402). Seoul, Republic of Korea: ACM Press. DOI: https://doi.org/10.1145/2702123.2702252

Buschek, D., De Luca, A., & **Alt, F.** (2016). Evaluating the influence of targets and hand postures on touch-based behavioural biometrics. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems – CHI '16* (pp. 1349–1361). Santa Clara, California, USA: ACM Press. DOI: https://doi.org/10.1145/2858036.2858165

Corallo, G., Sackur, J., Dehaene, S., & **Sigman, M.** (2008). Limits on introspection: Distorted subjective time during the dual-task bottleneck. *Psychological Science, 19*, 1110 –1117. DOI: https://doi.org/10.1111/j.1467-9280.2008.02211.x

Dedman, B. (2008, April 9). New anti-terror weapon: Hand-held lie detector. *NBC News*. Retrieved from http://www.nbcnews.com/id/23926278/ns/world_news-terrorism/t/new-anti-terror-weapon-hand-held-lie-detector/

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. DOI: https://doi.org/10.5334/irsp.82

Eisinga, R., Grotenhuis, M. te, & **Pelzer, B.** (2013). The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown? *International Journal of Public Health, 58*(4), 637–642. DOI: https://doi.org/10.1007/s00038-012-0416-3

Foss, D. J. (1982). A discourse on semantic priming. *Cognitive Psychology, 14*(4), 590–607. DOI: https://doi.org/10.1016/0010-0285(82)90020-2

Foster, E. D., & **Deardorff, A.** (2017). Open Science Framework (OSF). *Journal of the Medical Library Association*, *105*(2). DOI: https://doi.org/10.5195/JMLA.2017.88

Goel, M., Jansen, A., Mandel, T., Patel, S. N., & **Wobbrock, J. O.** (2013). ContextType: Using hand posture information to improve mobile touch screen text entry. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI '13* (p. 2795). Paris, France: ACM Press. DOI: https://doi.org/10.1145/2470654.2481386

Gordon, N. J. (2017). *Essentials of polygraph and polygraph testing*. Boca Raton, FL: CRC Press, Taylor & Francis Group. DOI: https://doi.org/10.1201/9781315438641

Hartwig, M., & **Bond, C. F.** (2011). Why do lie-catchers fail? A lens model meta-analysis of human lie judgments. *Psychological Bulletin*, *137*(4), 643–659. DOI: https://doi.org/10.1037/a0023589

Jarosz, A. F., & **Wiley, J.** (2014). What are the odds? A practical guide to computing and reporting Bayes factors. *The Journal of Problem Solving*, *7*(1), 2–9. DOI: https://doi.org/10.7771/1932-6246.1167

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, England: Clarendon Press.

Kay, M., Rector, K., Consolvo, S., Greenstein, B., Wobbrock, J., Watson, N., & **Kientz, J.** (2013). PVT-Touch: Adapting a reaction time test for touchscreen devices. In *Proceedings of the ICTs for improving Patients Rehabilitation Research Techniques*. Venice, Italy: IEEE. DOI: https://doi.org/10.4108/icst.pervasivehealth.2013.252078

Kelley, K. (2019). MBESS: The MBESS R Package. *R package version 4.5.1*. Retrieved from https://CRAN.R-project.org/package=MBESS

Khandeparkar, A., Gupta, R., & **Sindhya, B.** (2015). An introduction to hybrid platform mobile application development. *International Journal of Computer Applications*, *118*(15), 31–33. DOI: https://doi.org/10.5120/20824-3463

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

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. DOI: https://doi.org/10.1016/j.jarmac.2015.11.004

Krapohl, D. J. (2011). Limitations of the Concealed Information Test in criminal cases. In B. Verschuere, G. Ben-Shakhar & E. Meijer (Eds.), *Memory Detection* (pp. 151–170). Cambridge, England: Cambridge University Press. Retrieved from http://ebooks.

cambridge.org/ref/id/CBO9780511975196A022. DOI: https://doi.org/10.1017/CBO9780511975196.009

Kraut, R. (1980). Humans as lie detectors. *Journal of Communication*, *30*(4), 209–218. DOI: https://doi.org/10.1111/j.1460-2466.1980.tb02030.x

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. DOI: https://doi.org/10.3389/fpsyg.2013.00863

Lakens, D. (2017). Equivalence tests: A practical primer for *t* tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*(4), 355–362. DOI: https://doi.org/10.1177/1948550617697177

Lakens, D., Schneider, I. K., Jostmann, N. B., & **Schubert, T. W.** (2011). Telling things apart: The distance between response keys influences categorization times. *Psychological Science*, *22*(7), 887–890. DOI: https://doi.org/10.1177/0956797611412391

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

Lehmann, F., & **Kipp, M.** (2018). How to hold your phone when tapping: A comparative study of performance, precision, and errors. In *Proceedings of the 2018 ACM International Conference on Interactive Surfaces and Spaces – ISS '18* (pp. 115–127). Tokyo, Japan: ACM Press. DOI: https://doi.org/10.1145/3279778.3279791

Lin, Y.-C. (2013). The relationship between touchscreen sizes of smartphones and hand dimensions. In C. Stephanidis & M. Antona (Eds.), *Universal Access in Human-Computer Interaction. Applications and Services for Quality of Life* (Vol. 8011, pp. 643–650). Berlin, Germany: Springer Berlin Heidelberg. DOI: https://doi.org/10.1007/978-3-642-39194-1_74

Lukács, G. (2019). neatStats: An R Package for neat and painless statistical reporting. *R package version 0.3.1*. Retrieved from https://github.com/gasparl/neatstats

Lukács, G., & **Ansorge, U.** (2019a). Information leakage in the Response Time-Based Concealed Information Test. *Applied Cognitive Psychology*. DOI: https://doi.org/10.1002/acp.3565

Lukács, G., & **Ansorge, U.** (2019b). Methodological improvements of the association-based concealed information test. *Acta Psychologica*, *194*, 7–16. DOI: https://doi.org/10.1016/j.actpsy.2019.01.010

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

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. DOI: https://doi.org/10.1016/j.jarmac.2017.01.013

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

Makowski, D., Ben-Shachar, M., & **Lüdecke, D.** (2019). bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework. *Journal of Open Source Software, 4*(40), 1541. DOI: https://doi.org/10.21105/joss.01541

Meijer, E. H., Selle, N. K., Elber, L., & **Ben-Shakhar, G.** (2014). Memory detection with the Concealed Information Test: A meta analysis of skin conductance, respiration, heart rate, and P300 data: CIT meta-analysis of SCR, respiration, HR, and P300. *Psychophysiology*, *51*(9), 879–904. DOI: https://doi.org/10.1111/psyp.12239

Meixner, J. B., & **Rosenfeld, J. P.** (2011). A mock terrorism application of the P300-based concealed information test. *Psychophysiology, 48*(2), 149–154. DOI: https://doi.org/10.1111/j.1469-8986.2010.01050.x

Meyer, D. E., & **Schvaneveldt, R. W.** (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology, 90*(2), 227–234. DOI: https://doi.org/10.1037/h0031564

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

National Research Council. (2010). *Field evaluation in the intelligence and counterintelligence context: Workshop summary*. Washington, DC: National Academies Press. DOI: https://doi.org/10.17226/12854

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

Ogawa, T., Matsuda, I., Tsuneoka, M., & **Verschuere, B.** (2015). The concealed information test in the laboratory versus Japanese field practice: Bridging the scientist-practitioner gap. *Archives of Forensic Psychology, 1*(2), 16–27.

Podlesny, J. A. (2003). A paucity of operable case facts restricts applicability of the Guilty Knowledge Technique in FBI criminal polygraph examinations. *Forensic Science Communications, 5*(3).

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

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. DOI: https://doi.org/10.3758/s13428-014-0471-1

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. DOI: https://doi.org/10.1007/s10979-005-6832-7

Schatz, P., Ybarra, V., & **Leitner, D.** (2015). Validating the accuracy of reaction time assessment on computer-based tablet devices. *Assessment*, *22*(4), 405–410. DOI: https://doi.org/10.1177/1073191114566622

Schuirmann, D. J. (1987). A comparison of the Two One-Sided Tests Procedure and the Power Approach for assessing the equivalence of average bioavailability.

Journal of Pharmacokinetics and Biopharmaceutics, *15*(6), 657–680. DOI: https://doi.org/10.1007/BF01068419

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. DOI: https://doi.org/10.3758/CABN.9.1.71

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

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 1904–1920, 3*(3), 271–295. DOI: https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

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. DOI: https://doi.org/10.1037/bul0000087

United States Office of the Secretary of Defense. (2018). *Department of defense budget fiscal year (FY) 2019. Justification for FY 2019 Overseas Contingency Operations (OCO). Afghanistan Security Forces Fund (ASFF).* (No. 7–8119289). Retrieved from https://comptroller.defense.gov/Portals/45/Documents/defbudget/fy2019/FY2019_ASFF_Justification_Book.pdf

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. DOI: https://doi.org/10.1016/j.sbspro.2014.03.335

Verschuere, B., & **Kleinberg, B.** (2015). ID-check: Online Concealed Information Test reveals true identity. *Journal of Forensic Sciences*, *61*(S1), S237–S240. DOI: 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. DOI: https://doi.org/10.1016/j.jarmac.2015.01.001

Verwey, W. B. (1995). A forthcoming key press can be selected while earlier ones are executed. *Journal of Motor Behavior*, *27*(3), 275–284. DOI: https://doi.org/10.1080/00222895.1995.9941717

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,* 146. DOI: https://doi.org/10.3389/fpsyg.2013.00146

Vrij, A., & **Granhag, P. A.** (2012). Eliciting cues to deception and truth: What matters are the questions asked. *Journal of Applied Research in Memory and Cognition, 1*(2), 110–117. DOI: https://doi.org/10.1016/j.jarmac.2012.02.004

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. DOI: https://doi.org/10.3758/BF03194105

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & **Kievit, R. A.** (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*(6), 632–638. DOI: https://doi.org/10.1177/1745691612463078

Wang, F., & **Ren, X.** (2009). Empirical evaluation for finger input properties in multi-touch interaction. In *Proceedings of the 27th international conference on Human factors in computing systems – CHI '09* (p. 1063). Boston, MA: ACM Press. DOI: https://doi.org/10.1145/1518701.1518864

Wobbrock, J. O., Myers, B. A., & **Aung, H. H.** (2008). The performance of hand postures in front- and back-of-device interaction for mobile computing. *International Journal of Human-Computer Studies*, *66*(12), 857–875. DOI: https://doi.org/10.1016/j.ijhcs.2008.03.004

Zou, K. H., O'Malley, A. J., & **Mauri, L.** (2007). Receiver operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation*, *115*, 654–657. DOI: https://doi.org/10.1161/CIRCULATIONAHA.105.594929

Peer review comments
The author(s) of this paper chose the Open Review option, and the peer review comments can be downloaded at:
http://doi.org/10.1525/collabra.255.pr

How to cite this article: Lukács, G., Kleinberg, B., Kunzi, M., & Ansorge, U. (2020). Response Time Concealed Information Test on Smartphones. *Collabra: Psychology,* 6(1): 4. DOI: https://doi.org/10.1525/collabra.255

Senior Editor: Simine Vazire

Editor: Antonio Freitas

Submitted: 24 April 2019 **Accepted:** 09 December 2019 **Published:** 09 January 2020

UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Collabra: Psychology is a peer-reviewed open access journal published by University of California Press.

OPEN ACCESS