Journal of Applied Research in Memory and Cognition xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Journal of Applied Research in Memory and Cognition

journal homepage: www.elsevier.com/locate/jarmac

Familiarity-Related Fillers Improve the Validity of Reaction Time-Based Memory Detection☆

Gáspár Lukács, Bennett Kleinberg, and Bruno Verschuere*

University of Amsterdam, The Netherlands

The reaction time (RT)-based Concealed Information Test (CIT) allows for the detection of concealed knowledge (e.g., one's true identity) when the questions are presented randomly (multiple-probe protocol), but its performance is much weaker when questions are presented in blocks (e.g., first question about surname, then about birthday; single-probe protocol). The latter test protocol, however, is the preferred and sometimes even the only feasible interviewing method in real-life. In a first, preregistered, experiment ($n = 363$), we show that the validity of the single-probe protocol version can be substantially improved by including familiarity-related fillers: stimuli related to either familiarity (e.g., the word "FAMILIAR,") or unfamiliarity (e.g., the word "UNFAMILIAR"). We replicated these findings in a second, preregistered, experiment ($n = 237$), where we further found that the use of familiarity-related fillers even improved the classic multiple-probe protocol. We recommend the use of familiarity-related filler trials for the RT-based CIT.

General Audience Summary

The Concealed Information Test (CIT) assesses recognition of concealed information, for instance about a crime or about one's true identity. For this purpose, the CIT initially relied upon physiological responses recorded with a polygraph. Nowadays, administration can be done more easily, through the recording of reaction times. In two studies, participants tried to hide their own identity (i.e., respond UNKNOWN) and pretend to be someone else (i.e., respond KNOWN to a false identity). They also responded with UNKNOWN to unrelated identity items. Reaction times allowed the detection of someone's concealed identity. More importantly, we show that detection efficiency can be significantly improved through the addition of trials that require the classification of familiarity-related words (e.g., "known" and "unknown"). We see at least two possible reasons why the inclusion of such familiarity-related words helps: They assure processing of relevant item content, and they increase response conflict (by increasing reliance upon familiarity). In sum, familiarity-related words consistently increased the validity of reaction time-based concealed information detection.

Keywords: Memory detection, Reaction times, Lie detection, Concealed Information Test, Polygraph, Deception

Enhancing the RT-Based Concealed Information Test with Filler Items

Reliable and effective methods for deception detection are widely needed, for instance in criminal proceedings, in pre-employment screening, and for national security issues. The Concealed Information Test (CIT) aims to reveal whether a person is concealing knowledge regarding a certain detail (Lykken, 1959; for a review see Verschuere, Ben-Shakhar, & Meijer, 2011). To illustrate the CIT, let us consider a murder case scenario in which the murder weapon is known only to the perpetrator and the investigators. In this case, the CIT could include the actual murder weapon (the *probe*; e.g. "rifle") and several other weapons (*irrelevants*; e.g. "brick" "bat", "statue",

Author Note

☆ Please note that this paper was handled by the current editorial team of JARMAC.

* Correspondence concerning this article should be addressed to Bruno Verschuere, University of Amsterdam, Department of Clinical Psychology, The Netherlands. Contact: b.j.verschuere@uva.nl

Please cite this article in press as: Lukács, G., et al. Familiarity-Related Fillers Improve the Validity of Reaction Time-Based Memory Detection. *Journal of Applied Research in Memory and Cognition* (2017), http://dx.doi.org/10.1016/j.jarmac.2017.01.013

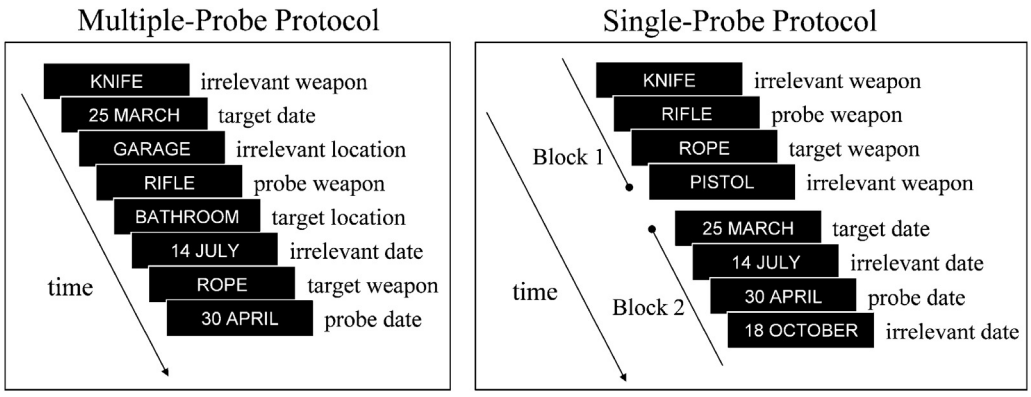

Figure 1. The multiple-probe and the single-probe protocols of the CIT, illustrated with a hypothetical murder case. In case of the multiple-probe protocol (left panel), all items in all categories are randomly intermixed throughout the task, while in case of the single-probe protocol (right panel), the items of each category are all presented within a separate block. Irrespective of protocol, each category involves one probe (e.g., in this case, *rifle* for actual murder weapon and *30 April* for actual date of the murder), one target (e.g., *rope* and *25 March*), and several, usually four irrelevant items (e.g., *knife* and *14 July*). (For brevity, the larger ratio of irrelevants is not proportionately represented in this illustration.)

and "knife") as the items that would be sequentially presented to a suspect in a random order. The recognition of the probe (in this case "rifle") by a guilty suspect will typically result in a stronger physiological response in comparison to the irrelevant items, which can be picked up by a polygraph.

Exploring the electrophysiological signature of concealed information, Farwell and Donchin (1991) found pronounced reaction time differences between concealed and irrelevant items. Seymour, Seifert, Shafto, and Mosmann (2000) showed that reaction time in itself can be a valid index of concealed information. The reaction time (RT)-based CIT is based merely on behavioral reactions: When an examinee needs to respond to each item with a keypress, the recognition of the probe will typically result in a slower response to that item than to the irrelevant items. Along with the probes and the irrelevants, the RT-based CIT includes *targets*, which requires a different response than the rest of the items. Originally, the inclusion of targets merely served to ensure attention to the stimuli. Targets may actually contribute to the validity of the RT-based CIT. Targets create a response conflict for the probes in knowledgeable participants. Fast, familiarity-related responding allows for the target (familiar) versus non-target (non-familiar; irrelevant) classification. For probes, however, familiarity-related responding leads to the wrong response (i.e., probes are familiar, but should be classified as being unfamiliar, like the irrelevant items), thereby slowing down the reaction time (Verschuere & De Houwer, 2011). This RT-based CIT is a very practical method, since it is short, only requires a regular computer, and can be easily standardized.

Different Test Protocols: The Single-Probe and the Multiple-Probe Protocol

In the classic, polygraph-based CIT, questions (about the murder weapon, about the locations, etc.) are asked one by one. This protocol has been called the single-probe protocol, where each category is presented in separate blocks (see right panel in Figure 1). The RT-based CIT, however, has typically relied on the multiple-probe protocol (see left panel in Figure 1), where items related to the different categories (e.g., weapons, locations, dates) are completely intermixed throughout the task.

Evidence suggests that, for the RT-based CIT, the single probe protocol leads to lower validity than the multiple probe protocol. One study did not find significant differences for RTs between the two protocols (Rosenfeld, Shue, & Singer, 2007), but the small sample size ($n = 9$–13 per condition) restricted statistical power and therefore does not allow firm conclusions with regard to protocol differences. In two studies ($n_{Study1} = 40$, $n_{Study2} = 210$) Verschuere, Kleinberg, and Theocharidou (2015) found that the single-probe protocol of RT-based CIT has significantly lower validity than the multiple-probe protocol. The results by Eom, Sohn, Park, Eum, and Sohn (2016) also point to this direction. In their mock crime study, the RT-based CIT was ineffective when testing for a single probe item, but highly effective when testing for 3–5 probe items.

The multiple-probe protocol is the preferred protocol from a validity perspective, but may hinder introduction of the RT-based CIT to real-life practice. In Japan, the only country using the CIT on a large scale, practitioners currently consider the single-probe protocol to be the only viable option (Ogawa, Matsuda, Tsuneoka, & Verschuere, 2015). This preference may partially be a simple convention, but there may also be solid reasons behind it. First, in real-life cases, the number of applicable probe items (details that only a guilty person would recognize) is often limited, and it may even happen that only a single item is available for a given examination (Podlesny, 2003). Clearly, when there is only one probe, the multiple-probe protocol cannot be used. Second, for some common test procedures and scoring algorithms of the CIT, only the single-probe protocol can be used. The Peak of Tension test (Krapohl, 2011) is a well-known example. One may ask the examinee, for instance, how many people were involved in the crime. The answer options may then be presented in ascending order (1–2–3–4–etc.). Practitioners have argued that such a fixed order may add to validity because the building tension toward the correct item and the subsequent relief may be used as signs of recognition. Third, another practical advantage of the single-probe protocol is the possibility of sequential testing. Based on the results of one

Please cite this article in press as: Lukács, G., et al. Familiarity-Related Fillers Improve the Validity of Reaction Time-Based Memory Detection. *Journal of Applied Research in Memory and Cognition* (2017), http://dx.doi.org/10.1016/j.jarmac.2017.01.013

test (using one probe), the practitioner may decide how to proceed with the following tests. This also allows the CIT to be used for an exploratory examination that sequentially narrows down the possibilities in a scenario when the probe is not known (Meixner & Rosenfeld, 2011). For example, to find the location of an upcoming terrorist attack using an RT-based CIT, a suspect would be presented with several countries where the attack is plausible. The country to which the slowest responses were made would be presumed to be the actual location of the upcoming attack. Subsequently, the suspect would be shown cities in that country. Then it could be further narrowed to street names, etc. The initial aim of our study was to improve the validity of the single-probe protocol RT-based CIT.

Improving the Validity of the RT-based CIT Through Filler Items

We reasoned that it could help to add familiarity-related filler trials that also needed to be categorized as familiar ("FAMILIAR," "RECOGNIZED," and "MINE") versus unfamiliar ("UNFAMILIAR," "UNKNOWN," "OTHER," "THEIRS," "THEM," and "FOREIGN"), through key presses, similar to the target versus non-target classification. Such familiarity-related fillers may benefit the single probe protocol for several reasons.

First, the filler items may assure semantic processing and increase the attention to the relevant dimension of CIT stimuli (familiarity or significance). The single-probe protocol allows participants to ignore the probes by simply focusing on those aspects of the target item that are necessary for the target/non-target discrimination, thereby effectively ignoring the familiarity and significance of the probes. An example may illustrate the problem: a single probe CIT for a date (probe: APRIL 2; target: JANUARY 13; irrelevant: MAY 6, SEPTEMBER 8, DECEMBER 21, FEBRUARY 1) may not be valid when the participant approaches the task in a way that allows to effectively ignore the meaning of the probe (e.g., a strategy such as "only press left for JANUARY", or even more extreme "only press left for stimuli starting with J"). Suchotzki, Verschuere, Crombez, and De Houwer (2013) have shown the crucial importance of the semantic processing of the items in an RT-based lie detection task: If participants categorized the items according to an irrelevant feature that does not relate to the relevant truth value of the item (e.g., the color of the sentences determining the YES or NO response), there were no lie–truth RT differences.

Second, such fillers may strengthen response conflict on probe trials. The inclusion of the filler trials stresses the target (familiar) versus non-target (non-familiar) discrimination task, thereby possibly strengthening familiarity-related responding to the targets. In sum, we saw several reasons why familiarity-related fillers may improve the validity of the RT-based CIT. The focus in the current paper is on establishing *whether* that is indeed the case, without trying to disentangle *why*.

The main purpose of Experiment 1 was to examine whether the validity of the single-protocol RT-based CIT could be improved through the use of familiarity-related filler items. Thus, the key comparison was between knowledgeable individuals tested using the standard single-probe protocol *without* familiarity-related fillers and those tested using single-probe protocol *with* familiarity-related fillers. We also included a condition using the *multiple-probe* protocol to examine the extent of the improvement: Can the single-probe protocol reach the validity of the multiple-probe protocol by adding fillers?

Experiment 1

The hypotheses, method, and analyses were preregistered before data collection: https://aspredicted.org/x7pww.pdf. All data are publically available at https://osf.io/kv65n/.

Method

Participants. This experiment was run on CrowdFlower (www.crowdflower.com), an online crowdsourcing platform where participants from anywhere in the world can register to complete small online tasks (see Peer, Samat, Brandimarte, & Acquisti, 2015). We opened slots for 500 participants for our experiment, paying 1.50 USD per completed task. Only those contributors who had a track record for accurate performance on CrowdFlower tasks ('level III contributors') were allowed to participate. CrowdFlower serves as a gateway for several websites (e.g., inboxdollars.com), and tracks the proportion of trusted judgments. We excluded websites with more than 20% untrusted judgments.

Each participant was randomly assigned to perform the CIT in one of the three protocols: single-probe without fillers, single-probe with familiarity-related fillers, or multiple-probe without fillers. Each participant was also randomly assigned to the *naive* or *knowledgeable* condition. In the knowledgeable condition, the probe items were participants' self-reported autobiographical identity details (e.g., their country of origin), simulating a guilty suspect. In the naive condition, the probe items were not the identity details of the participants (simulating an innocent suspect). In this latter case, additional randomly selected irrelevant items served as probe items. Note that the naive conditions were not strictly needed to test our hypotheses, but are of use to calculate individual detection accuracy (ROCs, see below).

Due to simultaneous starting times, not 500, but 502 participants completed the task. For nine participants, the recorded data were incomplete, probably due to not following correctly the technical procedure (e.g., changed the URL during task) or due to connection problems. We applied exclusion criteria in three steps (see Kleinberg & Verschuere, 2015). First, to ensure that participants did not participate more than once, all data coupled to an IP address that was recorded more than once were deleted, resulting in 91 exclusions. Second, we excluded 36 participants[1] who had 50% or more errors on any of the trial

[1] We sought to examine whether the addition of familiarity-related filler trials affected the RT-CIT as it is commonly applied. Following Kleinberg and Verschuere (2016), Verschuere and Kleinberg (2015), and Verschuere et al. (2015), the participants in the RT-CIT without fillers were explained that their task was to indicate recognition versus non-recognition. Such a recognition judgment, however, is not meaningful for the fillers (without further explanation, participants would probably classify all fillers as non-recognized as they have not been previously presented). Therefore, we changed the task so that it

Please cite this article in press as: Lukács, G., et al. Familiarity-Related Fillers Improve the Validity of Reaction Time-Based Memory Detection. *Journal of Applied Research in Memory and Cognition* (2017), http://dx.doi.org/10.1016/j.jarmac.2017.01.013

Table 1
Dropout and Exclusion Rates in Experiment 1

	Knowledgeable			Naive		
	Single-probe without fillers *n* (% original *n*)	Single-probe with fillers *n* (% original *n*)	Multiple-probe without fillers *n* (% original *n*)	Single-probe without fillers *n* (% original *n*)	Single-probe with fillers *n* (% original *n*)	Multiple-probe without fillers *n* (% original *n*)
Initial *n*	107	74	101	90	66	91
Dropout	1	10	7	6	6	5
Exclusions	21	29	23	13	21	24
Final *n*	85 (79%)	35 (47%)	71 (70%)	71 (79%)	39 (59%)	62 (68%)

Note. Initial number of participants who began the test is an estimate: it is the number of participants who finished the first practice task. Dropout is the number of participants who did not complete the whole task. Exclusions is the number of participants that we excluded.

types: (a) probes, (b) targets, (c) irrelevants, and (in the case of the version with filler trials) (d) familiarity-related fillers and (e) unfamiliarity-related fillers. Third, we excluded three participants due to having fewer than 50% remaining trials for RT analysis (correct responses with an RT between 150 and 800 ms). Detailed dropout and exclusion rates for all conditions are shown in Table 1.

Thus, the final sample consisted of 363 participants, who had been randomly assigned to one of the six conditions: 85 in the simple single-probe knowledgeable condition (age $M = 33.82$ years, $SD = 9.15$; 64.71% male), 35 in the enhanced single-probe knowledgeable condition (age $M = 34.63$ years, $SD = 8.70$; 62.86% male), 71 in the simple multiple-probe knowledgeable condition (age $M = 35.51$ years, $SD = 9.91$; 73.24% male), 71 in the simple single-probe naive condition (age $M = 33.80$ years, $SD = 9.32$; 63.38% male), 39 in the enhanced single-probe naive condition (age $M = 33.44$ years, $SD = 8.53$; 58.97% male), and 62 in the simple multiple-probe naive condition (age $M = 34.35$ years, $SD = 8.87$; 64.52% male). There were no significant differences across conditions in respect of age ($F(5, 357) = 0.41$, $p = .842$, $f = 0.03$) or gender ($\chi^2(5) = 2.90$, $p = .715$, Cramer's $V = 0.04$).

Procedure. The online experimental task was written in HTML5/JavaScript framework (see Kleinberg & Verschuere, 2015, 2016; Verschuere & Kleinberg, 2015; Verschuere et al., 2015), and the entire task is available via this link: http://liereseach.com/wp-content/uploads/onepmp/exp1/intro/gen_intro2.html. The data for both experiments can be retrieved from the Open Science Framework data repository via https://osf.io/kv65n/ (Open Science Collaboration, 2012).

Upon accessing the link, participants gave informed consent in order to proceed further. Participants then provided demographic information (age, gender, mother tongue, and education) and chose, from a dropdown menu, the four autobiographical details that were subsequently used as probes in the CIT task: country of origin, date of birth (day and month), favorite color, and favorite animal.

This was followed by the RT-based CIT (see below). After the CIT, there was a short survey (see Appendix A at https://osf.io/kv65n/) and then participants received automatic individual feedback about whether they successfully avoided detection, and were given a brief explanation about the purpose of the study.

Reaction time-based concealed information test. Participants were informed that the RT-CIT simulates a lie detection scenario, during which they should try to hide their identities.

Item selection. Participants were then presented a short list of items within each of the four categories in the task (countries, dates, colors, and animals). The items on this list never contained any of the actual identity details of the given participants. The participants were asked to choose any (but a maximum of two per category) item that were personally meaningful to them or in any way appeared different from the rest of the items on those lists. Subsequently, unbeknownst to the participants, the items for the task were randomly selected from the non-chosen items (as this assures that the irrelevant items were indeed irrelevant). For a participant in the naive condition, six items were selected for each of the four categories and were randomly assigned to be one of the three item types: one probe, one target, or one of the four irrelevant items. For a participant in the knowledgeable condition, items were randomly selected to be the target or one of the four irrelevant items, and their self-reported identity details served as the probe item in each of the four categories. Thus, in either condition, there were altogether 24 unique items: four probes, four targets, and 16 irrelevant items.

Target learning. Next, participants were presented the four target items, and were asked to memorize these items in order to pretend to recognize them as their own 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.

RT-CIT. In the RT-CIT, the items were presented one by one in the center of the screen, and participants had to categorize them by pressing one of two keys ("e" or "i") on their keyboard.

is readily obvious for participants how to classify the fillers. Specifically, the participants' task was to categorize stimuli are being familiar versus unfamiliar, thereby making it obvious that familiarity-related fillers (e.g., "FAMILIAR") require assignment to the familiar category and unfamiliarity-related fillers (e.g., "UNFAMILIAR") require assignment to the non-familiar category. We later realized that this raises a possible confound because the framing of the tasks differs, and simply having the RT-CIT framed in terms of familiarity rather than recognition may have affected validity. We address this potential confound in Experiment 3, in which both the RT-CIT with fillers and the RT-CIT without fillers were framed in terms of familiarity.

Please cite this article in press as: Lukács, G., et al. Familiarity-Related Fillers Improve the Validity of Reaction Time-Based Memory Detection. *Journal of Applied Research in Memory and Cognition* (2017), http://dx.doi.org/10.1016/j.jarmac.2017.01.013

The design of the RT-CIT without fillers (either single-probe or multiple-probe protocol) replicated the regular RT-based CIT (e.g., Kleinberg & Verschuere, 2015, 2016; Verschuere & Kleinberg, 2015; Verschuere et al., 2015). Participants were told that pushing the "e" key means "YES," they recognize the item, while pushing the "i" key means "NO," they do not recognize the item, and they were correspondingly instructed to say YES to the details of their assumed identity that they were supposed to recognize as their own (the information that they had memorized on the previous screen, i.e. the targets), and NO to all other words (including those related to their true identity; i.e., both the irrelevants and the probes). As reminder the following captions remained displayed through the task: "Recognize?" on the top of the screen, "YES = e" on the left side, "NO = i" on the right side. In case of the RT-CIT with filler items, the description was slightly modified to focus on familiarity: participants were told that pushing the "e" key means that the displayed item is "FAMILIAR" to them, while pushing the "i" key means that the item is "UNFAMILIAR" to them. They were instructed to respond FAMILIAR to the details of their assumed identity (targets), and UNFAMILIAR to all other words (irrelevants and probes). In addition, they also had to categorize the filler items: those that referred to familiarity (e.g., "FAMILIAR" or "RECOGNIZED") had to be categorized as familiar ("e" key), while those that referred to unfamiliarity (e.g. "UNFAMILIAR" or "UNKNOWN") had to be categorized as unfamiliar ("i" key). The reminder captions in this case were: "Familiar to you?" on the top of the screen, "FAMILIAR = e" on the left side, "UNFAMILIAR = i" on the right side (see footnote 1).

Before the main task, there were three practice phases that increased in difficulty step-wise, so that the participants would gradually get used to the requirements of the main task (see Kleinberg & Verschuere, 2015). In each practice task, each of the four probe, four target, and 16 irrelevant items were presented once (thus altogether 24 stimuli). In the version with fillers, twelve filler items (four familiarity-related, eight unfamiliarity-related) were added to the task (thus altogether 36 stimuli, a third of which were filler items). The inter-stimulus interval between any two trials (i.e., between the end of one trial and the beginning of the next) was randomly chosen as either 250, 500, or 750 ms. The first practice task did not include a time limit, allowing participants to take as much time to respond as they needed, and each stimulus remained displayed until one of the two allowed response keys were pressed. In case of a correct response, the next trial followed. In case of an incorrect response, the word "WRONG" appeared below the stimulus in red color for 200 ms, followed by the next trial. In case of too few correct responses (below 50%) or too many suspiciously fast RTs (more than 20% faster than 150 ms), the participant received a corresponding feedback, was reminded of the instructions, and had to repeat the practice task. In the second practice phase, a response deadline was introduced: in case of no response within 1500 ms (counting from the appearance of the stimulus), the next trial followed. In addition to the previous criteria (as in the first practice task), participants also had to repeat this practice task if their average RT was too slow (beyond 800 ms). In the third practice phase, the participants received feedback for slow responses: if there was no response within 800 ms, the message "TOO SLOW" appeared in red color above the stimulus for 200 ms (but the stimulus remained displayed until a keypress or until 1500 ms passed). The criteria for passing or repeating this task were the same as in the second practice task.

The main task had the same parameters as the third practice phase, except that it contained 18 times as many trials: 432 in case of the versions without fillers (72 probes, 72 targets, 288 irrelevant items), and 648 in case of the version with filler (72 probes, 72 targets, 288 irrelevant items, 72 familiarity-related fillers, 144 unfamiliarity-related fillers).

In the single-probe without fillers version, the items of each category were presented within one block. For example, colors in the first block, countries in the second, animals in the third, and dates in the fourth. The order of the categories was random, but with the restriction that low-salient and high-salient category blocks alternated. Within each of the four blocks, the order of the items was randomized in groups of six: first, all six items (one probe, one target, four irrelevant) in the given category were presented in a random order, then the same six items were presented in another random order. This procedure was repeated until each item was presented 18 times (thus $6 \times 18 = 108$ items in one block). There were no breaks between blocks.

In the multiple-probe without fillers version, all items in the task were presented intermixed, irrespective of category. The order of the items was randomized in groups of 24: first, all 24 items (four probe, four target, 16 irrelevant) were presented in a random order, then all 24 items were again presented in another random order. Again, this procedure was repeated until each item was presented 18 times (thus $24 \times 18 = 432$ items in the entire task).

In the single-probe with fillers version, the order of probe, target, and irrelevant items was randomized the same way as in the single-probe without fillers version. Filler items were subsequently 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 (three familiarity-related, six unfamiliarity-related) preceded each of the four probes, four targets, and 16 irrelevant items exactly one time. (Thus $9 \times 24 = 216$ filler items were presented, and 216 out of the 432 other items were preceded by a filler item.)

Analysis plan. For all RT analyses, only trials with correct responses and with an RT between but inclusive of 150 and 800 ms were used. For follow-up contrasts, we report Cohen's d effect sizes using the following formula. For within-subject contrasts, $d_{within} = M(\mathrm{RT}_{(\mathrm{probes})} - \mathrm{RT}_{(\mathrm{irrelevants})}/\sqrt{(SD_{(\mathrm{probes})^2} + SD_{(\mathrm{irrelevants})^2} - 2 * r * SD_{\mathrm{probes}} * SD_{(\mathrm{irrelevants})}}$, where r is the correlation between $\mathrm{RT}_{(\mathrm{probes})}$ and $\mathrm{RT}_{(\mathrm{irrelevants})}$. For between-subjects contrasts, $d_{\mathrm{between}} = (M_{\mathrm{RT(Probe\text{-}Irrelevant\ Difference\ knowledgeable)}} - M_{\mathrm{RT(Probe\text{-}Irrelevant\ Difference\ naive)}})/\sqrt{(((n_{\mathrm{knowledgeable}} - 1) * SD_{(\mathrm{Probe\text{-}Irrelevant\ Difference\ knowledgeable})^2} + (n_{\mathrm{naive}} - 1) * SD_{(\mathrm{Probe\text{-}Irrelevant\ Difference\ naive})^2})/n_{\mathrm{knowledgeable}} + n_{\mathrm{naive}} - 2)}$ (see Kleinberg & Verschuere, 2015, 2016; Verschuere & Kleinberg, 2015; Verschuere et al., 2015; adopted from Lakens, 2013). We used an alpha level of .05 for all statistical tests.

Please cite this article in press as: Lukács, G., et al. Familiarity-Related Fillers Improve the Validity of Reaction Time-Based Memory Detection. *Journal of Applied Research in Memory and Cognition* (2017), http://dx.doi.org/10.1016/j.jarmac.2017.01.013

Results

Means RT (with SDs) for the probes and irrelevant items in naive and knowledgeable conditions for each protocol are given in Table 2, along with effect sizes (d_{within}) for the differences of probe mean RTs and irrelevant mean RTs.[2]

A 2 (Condition: knowledgeable vs. naive) by 3 (CIT protocol: (single-probe without fillers, single-probe with fillers, multiple-probe without fillers) between-subjects ANOVA on probe-irrelevant RT differences[3] revealed a significant main effect for Condition, $F(1, 357) = 158.47$, $p < .001$, $f = 0.67$, a significant main effect of CIT protocol, $F(2, 357) = 27.75$, $p < .001$, $f = 0.27$, and a significant interaction between Condition and CIT protocol, $F(2, 357) = 16.70$, $p < .001$, $f = 0.21$.

Follow-up t-tests revealed that for knowledgeable participants, the probe-irrelevant RT difference in the single-probe with fillers CIT protocol ($M = 38.54$, $SD = 25.41$) was larger than in the single-probe without fillers CIT protocol ($M = 8.29$, $SD = 16.85$; $t(118) = 7.64$, $p < .001$, $d_{between} = 1.54$ [1.09, 1.97]). Furthermore, the probe-irrelevant RT difference in the multiple-probe without fillers CIT protocol ($M = 26.66$, $SD = 20.49$) was larger than that in the single-probe without fillers CIT protocol, $t(154) = 6.15$, $p < .001$, $d_{between} = 0.99$ [0.65, 1.32]). Finally, the probe-irrelevant RT difference in the single-probe with fillers CIT protocol was larger than for the multiple-probe without fillers CIT protocol, $t(104) = 2.59$, $p = .011$, $d_{between} = 0.53$ [0.12, 0.94].

For naive participants, as expected, there was no significant difference between the probe-irrelevant differences between any two of the three protocols (p's > .05).

Individual classification. To examine the accuracy of concealed information detection at the individual level (i.e., the efficiency of discriminating knowledgeable participants from naive participants), we calculated areas under the receiver operating characteristic curve (AUCs; see National Research Council, 2003, pp. 342–344). Using a given predictor variable (here the standardized probe-irrelevant RT differences), the AUC plots the true positive rates (ratio of knowledgeable participants correctly classified as knowledgeable) and false positive rates (ratio of naive participants incorrectly classified as knowledgeable) across all possible cut-off points, and gives an averaged value that can range from 0 to 1. An AUC of .5 implies chance level classification, while an AUC of 1 is indicative of perfect classification (when all guilty and innocent classifications can be correctly made based on the predictor values). As predictor variable for the AUCs, we used individual effect sizes of the probe-irrelevant differences d_{CIT}. This d_{CIT} was calculated with the formula introduced by Noordraven and Verschuere (2013): $(M_{RT\ (probes)} - M_{RT\ (irrelevants)})/SD_{RT\ (irrelevants)}$.

AUCs for each of the three protocols are displayed in Table 2. Diagnostic accuracy was modest for the single-probe without fillers CIT protocol, and very good for both the single-probe with fillers CIT protocol and the multiple-probe CIT protocol. Using DeLong's test for the statistical comparison of two AUC values (DeLong, DeLong, & Clarke-Pearson, 1988; Robin et al., 2011), we found that the AUC for the single-probe without fillers CIT protocol was significantly lower than that of the multiple-probe without fillers CIT protocol, $D(282.47) = 9.40$, $p < .001$. Likewise, the AUC value for the single-probe with fillers CIT protocol was larger than that of the single-probe without fillers CIT protocol, $D(224.82) = 12.34$, $p < .001$, and larger than that of the multiple-probe without fillers CIT protocol, $D(204.99) = 2.27$, $p = .024$.

The split-half Spearman–Brown odd-even reliability (Spearman, 1910; Brown, 1910) of the d_{CIT} after all reported exclusions was (for knowledgeable participants) was high for the single-probe with fillers CIT protocol, $\rho = .79$ [.62, .89] and the multiple-probe without fillers CIT protocol $\rho = .70$ [.56, .80]. Reliability was very poor for the single-probe without fillers CIT protocol, $\rho = .15$ [−.07, .35].[4]

Discussion

The single-probe without fillers protocol had lower reliability and validity than the multiple-probe protocol, replicating Verschuere et al. (2015). More importantly, we found that familiarity-related filler items significantly improved the validity of the single-probe protocol version of the RT-based CIT. In fact, it improved to the extent that it outperformed the current gold standard for RT-based tests (the multiple-probe protocol).

Experiment 2 had two main purposes. The first was to replicate the key finding of Experiment 1, that the familiarity-related filler items can improve the validity of the single probe protocol, now sampling from a different population through another online crowdsourcing platform. Second, inspired by the extent to which fillers added to validity, we examined whether the familiarity-related filler items might not just improve suboptimal test protocols, but might also benefit the optimal, multiple-probe protocol. We therefore added a multiple-probe protocol condition for which we included familiarity-related filler items.

Experiment 2

The hypotheses, method, and analyses were again preregistered before data analysis: https://aspredicted.org/dmnh8.pdf. All data are publically available and can be found at https://osf.io/kv65n/.

Method

Participants. This experiment was run on Prolific (www.prolific.ac), an online crowdsourcing platform. This platform provided a new population for our study sample, and there are indications in the literature that it provides higher quality data compared to CrowdFlower (Peer et al., 2015). We opened

[2] Error rates are reported in the online supplement at https://osf.io/kv65n/.

[3] This slightly differs from the preregistered analyses in which we included probe versus irrelevant as a factor instead. Those analyses revealed the same results and for sake of brevity and consistency with Exp2 we chose to drop probe versus irrelevant as a factor and use the difference score instead.

[4] For naive participants: $\rho = .22$ [−.10, .50] for the single-probe with fillers CIT protocol; $\rho = .05$ [−.19, .28] for the single-probe without fillers CIT protocol; and $\rho = .28$ [.03, .50] for the multiple-probe without fillers CIT protocol.

Please cite this article in press as: Lukács, G., et al. Familiarity-Related Fillers Improve the Validity of Reaction Time-Based Memory Detection. *Journal of Applied Research in Memory and Cognition* (2017), http://dx.doi.org/10.1016/j.jarmac.2017.01.013

Table 2
Means Reaction Times (in ms; with SDs), and Effect Size (Cohen's d) and Classification Accuracy (AUC) in Experiment 1

	Knowledgeable			Naive			AUC [CI]	d_{betwen}
	Probe	Irrelevant	d_{within}	Probe	Irrelevant	d_{within}		
Single-probe without fillers	421 (37)	413 (35)	0.49	419 (42)	422 (39)	−0.15	.68 [.59, .76]	0.66
Single-probe with fillers	517 (53)	479 (51)	1.52	488 (37)	488 (35)	0.02	.94 [.89, .99]	1.97
Multiple-probe without fillers	523 (43)	496 (42)	1.30	490 (37)	487 (37)	0.19	.84 [.77, .91]	1.33

slots for 300 participants, paying 2.50 GBP, to complete our experiment. Only those contributors were allowed to participate who had an approval rate of not lower than 80%, who were older than 18 and younger than 65, and who did not participate in the small pilot and feasibilities studies for the current study.

Each new participant was randomly assigned to perform the CIT in one of the four protocols: single-probe without fillers, single-probe with fillers, multiple-probe without fillers, and multiple-probe with fillers. In Experiment 1, the assignment of participants to conditions was completely random, and consequently the group sizes were somewhat disproportionate. To obtain more proportionate sample sizes in each of the conditions, we implemented a semi-random assignment in this second experiment: a new participant was always randomly assigned to one of those conditions in which, at that time point, fewer participants completed the task than in the other conditions (except when all conditions are completed by the same number of participants, in which case the assignment was again completely random).

Naive conditions were included in Experiment 1 in order to calculate ROCs; however, no probe-irrelevant differences are in fact expected between protocols. As a simplification of the study design that already includes four conditions (i.e., four CIT protocols), no naive conditions were included in Experiment 2. Instead, using established procedures (Meijer, Smulders, Johnston, & Merckelbach, 2007), we simulated naive conditions for ROC calculations. As in Experiment 1, the probe items were self-reported autobiographical identity details.

Due to simultaneous starting times, 305 participants completed the task. For one participant, the recorded data was incomplete, probably due to not following the technical procedure correctly or due to connection problems. As in Experiment 1, we applied exclusion criteria in three steps (see Kleinberg & Verschuere, 2015). First, all data were excluded with an IP address that was recorded more than once, resulting in 24 exclusions. Second, we excluded 40 participants who had 50% or more errors on any of the trial types: (a) probes, (b) targets, (c) irrelevants, and (in case of the version with filler trials) (d) familiarity-related fillers, and (e) unfamiliarity-related fillers. Third, we excluded three participants due to having fewer than 50% remaining trials for RT analysis (correct responses with an RT between 150 and 800 ms). Detailed dropout and exclusion rates for all conditions are shown in Table 3.

Thus, the final sample consisted of 237 participants: 54 in the single-probe without fillers condition (age $M = 29.09$ years, $SD = 9.77$; 62.96% male), 71 in the single-probe with fillers condition (age $M = 32.10$ years, $SD = 11.36$; 56.34% male), 63 in the multiple-probe without fillers condition (age $M = 30.87$ years, $SD = 10.39$; 60.32% male), and 49 in the multiple-probe with fillers condition (age $M = 32.00$ years, $SD = 9.62$; 58.18% male). There were no significant differences across conditions in respect of age ($F(3, 213) = 0.75$, $p = .526$, $f = 0.07$) or gender ($\chi^2(3) = 0.58$, $p = .900$, Cramer's $V = 0.03$).

Procedure. The same online experimental task was used as in Experiment 1 (the entire task for Experiment 2 is available at http://lieresearch.com/wp-content/uploads/onepmp/exp2/intro/gen_intro2.html). The procedure was also the same as in Experiment 1, except that the multiple-probe with fillers protocol was added as a fourth CIT version and that all four CIT versions were performed only in knowledgeable condition.

The multiple-probe with fillers protocol was based on the multiple-probe without fillers protocol (see Experiment 1): all items in the task were intermixed, irrespective of category, and the order of the items was randomized in groups of 24 (four probe, four target, 16 irrelevant; altogether $18 \times 24 = 432$ probe, irrelevant, and target items in the task). Filler items were subsequently placed among these items in a random order, but with the same restrictions as in Experiment 1. (Namely, a filler trial was never followed by another filler trial, and each of the nine fillers

Table 3
Dropout and Exclusion Rates in Experiment 2

	Single-probe without fillers *n* (% original *n*)	Single-probe with fillers *n* (% original *n*)	Multiple-probe without fillers *n* (% original *n*)	Multiple-probe with fillers *n* (% original *n*)
Initial *n*	70	110	75	85
Dropout	6	11	4	14
Exclusions	10	28	8	22
Final *n*	54 (77%)	71 (64%)	63 (84%)	49 (58%)

Please cite this article in press as: Lukács, G., et al. Familiarity-Related Fillers Improve the Validity of Reaction Time-Based Memory Detection. *Journal of Applied Research in Memory and Cognition* (2017), http://dx.doi.org/10.1016/j.jarmac.2017.01.013

Table 4
Means Reaction Times (in ms; with SDs), and Effect Size (Cohen's d) and Classification Accuracy (Simulated AUC) in Experiment 2

	Knowledgeable			AUC sim. [CI]
	Probe *M* (*SD*)	Irrelevant *M* (*SD*)	d_{within}	
Single-probe without fillers	426 (42)	410 (37)	0.95 [0.63, 1.27]	.83 [.75, .91]
Single-probe with fillers	519 (46)	483 (42)	1.77 [1.39, 2.14]	.95 [.91, .98]
Multiple-probe without fillers	510 (41)	485 (40)	1.15 [0.83, 1.47]	.74 [.64, .83]
Multiple-probe with fillers	543 (48)	501 (38)	1.73 [1.28, 2.17]	.94 [.90, 0.98]

preceded each of the four probes, four targets, and 16 irrelevant items exactly one time. Thus $9 \times 24 = 216$ filler items were presented, and 216 out of the 432 other items were preceded by a filler item.)

Results

Main RT analysis. Means and SDs of individual RT means for the probes and irrelevant items in all conditions are given in Table 4, along with effect sizes for the differences of probe mean RTs and irrelevant mean RTs (d_{within}).

We conducted a one-way ANOVA with the between-subjects factor CIT protocol (single-probe without fillers, single-probe with fillers, multiple-probe without fillers, and multiple-probe with fillers). The ANOVA revealed a significant effect of CIT protocol, $F(3, 233) = 16.39$, $p < .001$, $f = 0.25$.

Follow-up *t*-tests showed that the probe-irrelevant RT differences for the single-probe with fillers condition were larger than in the single-probe without fillers condition, $t(123) = 5.95$, $p < .001$, $d_{between} = 1.07$ [0.69, 1.45], but also larger than in the multiple-probe without fillers condition, $t(132) = 2.83$, $p = .005$, $d_{between} = 0.49$ [0.14; 0.83]. There was no significant difference between single-probe with fillers protocol and multiple-probe with fillers protocol, $t(118) = 1.44$, $p = .151$, $d_{between} = 0.27$ [0.10, 0.63]). The probe-irrelevant RT difference for the multiple-probe without fillers condition was larger than that in the single-probe without fillers condition, $t(115) = 2.66$, $p = .009$, $d_{between} = 0.49$ [0.12, 0.86]. Finally, the probe-irrelevant RT differences in the multiple-probe with fillers protocol were larger than those in the multiple-probe without fillers protocol, $t(110) = 3.70$, $p < .001$, $d_{between} = 0.70$ [0.32, 1.09].

Individual classification. For sake of completeness, we also conducted ROC analyses with simulated innocent participants (see Meijer et al., 2007). The simulation entailed the following 3 steps: (a) we took the mean and standard deviation of the d_{CIT} scores for the single-probe without fillers protocol, the single-probe with fillers protocol and the multiple-probe without fillers protocol; (b) based on these values, we simulated a normal distribution of the d_{CIT} scores with 172 observations (the number of naive participants in Experiment 1) equally distributed across the four CIT protocol conditions (i.e., 43 per condition); and (c) this simulation procedure was repeated for each participant and we randomly sampled one value per participant from the 172 simulated d_{CIT} values. The ROC analyses show that diagnostic accuracy was modest to good in the conditions without fillers and very high for the conditions with fillers (see Table 2). Using DeLong's test for the statistical comparison of two AUC values, we found that the AUC for the single-probe with fillers CIT protocol was higher than that of the single-probe without fillers CIT protocol, $D(135.96) = 2.62$, $p = .001$, and higher than that of the multiple without fillers protocol, $D(145.17) = 2.45$, $p = .016$. Further, there was no difference between the AUCs for the single-probe without fillers CIT protocol and the multiple-probe without fillers CIT protocol, $D(198.85) = -1.44$, $p = .151$, or between the AUC for the single-probe with fillers CIT protocol and the multiple-probe with fillers CIT protocol, $D(192.48) = -0.17$, $p = .867$.

The split-half Spearman–Brown odd-even reliability (Brown, 1910; Spearman, 1910) of the d_{CIT} was moderate to high for the single-probe with fillers condition, $\rho = .54$ [.35, .69], for the multiple-probe with fillers condition, $\rho = .75$ [.60, .85], and for the multiple-probe without fillers condition, $\rho = .61$ [.43, .75]. reliability was poor for the single-probe without fillers condition, $\rho = -.35$ [−.56, −.09].

General Discussion

Is it possible to detect concealed information based only upon the speed of a button press? This is an attractive idea, precisely because of its simplicity. The effect size of the RT-based CIT is typically large, but there is also large heterogeneity (Suchotzki, Verschuere, Van Bockstaele, Ben-Shakhar, & Crombez, 2017). In the present study we (a) confirm the large effect size, (b) confirm that test protocol moderates the effect size, and, most importantly, (c) show that a very simple intervention—the inclusion of familiarity-related fillers—significantly increased effect size. We now discuss these three observations.

First, at the most basic level, our findings align with recent qualitative reviews (Verschuere et al., 2015) and meta-analyses (Meijer, Verschuere, Gamer, Merckelbach, & Ben-Shakhar, 2016; Suchotzki et al., 2017) that the RT-based CIT typically results in a large effect size. In the present study, the size of the effect allowed detection of concealed identity details at the individual level. The area under the curve varied between .67 (suboptimal test protocol) and .95 (optimal test protocol), which implies that a randomly chosen knowledgeable examinee has a 67–95% chance to show a stronger response on the RT-based CIT than a randomly chosen naive examinee. Along with meta-analytic comparisons (Meijer et al., 2016) and direct

Please cite this article in press as: Lukács, G., et al. Familiarity-Related Fillers Improve the Validity of Reaction Time-Based Memory Detection. *Journal of Applied Research in Memory and Cognition* (2017), http://dx.doi.org/10.1016/j.jarmac.2017.01.013

experimental evidence (Verschuere, Crombez, Degrootte, & Rosseel, 2010), our data provide further support to the idea that RTs can reach the level of validity that have been obtained with established physiological measures.

Second, test protocol moderates effect size. Following up on the pioneering publications of the RT-based CIT (Farwell & Donchin, 1991; Seymour et al., 2000), most researchers have used the multiple-probe protocol, presenting items of all categories in a random order. Because of practitioner's preference (Ogawa et al., 2015) and dedicated test applications (e.g., sequential testing), our lab started working with the single-probe protocol, which combines the items of a category per block. Suddenly, the large effect sizes that we typically observed dropped dramatically. This impression was confirmed in two studies (Verschuere et al., 2015). With two new studies, we provide further evidence for the moderating role of test protocol. The multiple-probe protocol consistently outperformed the single-probe protocol in the RT-based CIT.

Third, while the single probe protocol is suboptimal, the use of familiarity-related fillers significantly improved its validity. In Experiment 1, the detection accuracy of the single-probe protocol increased from a modest AUC of .67 to an impressive AUC of .94 through the use of familiarity-related filler items. In Experiment 2 the AUC increased from .81 to again an impressive .96. To the best of our knowledge, this is the first *preregistered* lie detection study to obtain such a high diagnostic accuracy. With the study, code, and data being publically available, we opt for maximal transparency for others to control and repeat our findings. To our surprise, both experiments showed that the familiarity-related fillers improved the validity of the single probe protocol to the extent that it outperformed the gold standard, the multiple-probe protocol. Also, Experiment 2 showed that even the validity of the multiple-probe protocol can be improved through the inclusion of familiarity-related filler items. Our focus was on establishing whether, not why, the validity of the RT-based CIT would benefit from familiarity-related fillers. Follow-up research can explore possible mechanisms, and we have offered several options in the Introduction: hampering strategic focus on the target, deeper semantic processing, and greater familiarity-driven response conflict. Concurrent electrophysiological recordings, for instance, may help to shed light on the role of response conflict.

Limitations and Directions for Future Research

This study is not without its limitations. First, both experiments aimed at detecting concealed identity-related information, and it should be further explored whether our findings generalize to other applications of the CIT (e.g., to concealed crime information).

Second, our participant loss was substantial in both studies (16–53%). If this exclusion rate were true to real-life cases, it would severely limit the protocol's applicability. It should be noted that dropout may be related to low motivation (i.e., the small incentive for participation in this study), and the exclusions were set a priori to be very conservative. Participant loss was about twice as large when using fillers (36–53%) as compared to not using fillers (16–32%). Again, this may limit the protocol's applicability in real-life settings. Future research should explore whether participant loss can be reduced with more extensive practice and/or more motivated participants.

Third, apart from the use of fillers, there was another difference between test protocols: The task header either focused on recognition (protocols without fillers) or on familiarity (protocols with fillers). Therefore, the possibility exists that the improved validity obtained through using fillers is partly related to the focus on familiarity instead of recognition. We therefore ran a third, preregistered (http://aspredicted.org/4jn79.pdf), experiment ($n = 410$). Experiment 3 replicated the knowledgeable conditions of Experiment 1, now using the exact same familiarity header in all conditions. Experiment 3 confirmed again that the familiarity-related fillers significantly improved the validity of the single-probe protocol, to the level of that of the multiple probe protocol (for further details see https://osf.io/kv65n/).

Fourth, for a test that aims to assess concealed information, faking provides an important challenge—an issue unaddressed by the present study. While it has been argued that RTs would be particularly vulnerable to faking (Farwell & Donchin, 1991), empirical evidence is actually mixed. Seymour et al. (2000), for instance, found that the high accuracy of the RT-CIT was preserved despite informing participants about the rationale of the RT-CIT. The fast pace of the test and the response deadline may have prevented successful faking. Rosenfeld, Soskins, Bosh, and Ryan (2004), however, showed that with dedicated instructions on how to fake the test, the accuracy of the RT-CIT deteriorated significantly (see also Suchotzki et al., 2017). It is important to empirically assess faking vulnerability in future research.

Despite these limitations, our study also has important strengths, including the substantial sample size, the fact that we replicated the key effects, and that we preregistered our hypotheses, methods, and analytic plan.

Take Home Message

The practical implications of this study are straightforward: Adding familiarity-related fillers to the RT-based CIT helps to reveal whether a person is concealing knowledge about critical information. We are curious to see whether these effects generalize to other pieces of (e.g., crime-related) information, and to other (e.g., electrophysiological) response measures. The familiarity-related fillers make it possible to use protocols preferred by practitioners and thereby facilitate the implementation of the RT-based CIT in real life settings.

Conflict of Interest Statement

The authors declare no conflict of interest.

Please cite this article in press as: Lukács, G., et al. Familiarity-Related Fillers Improve the Validity of Reaction Time-Based Memory Detection. *Journal of Applied Research in Memory and Cognition* (2017), http://dx.doi.org/10.1016/j.jarmac.2017.01.013

Author Contributions

Designed the study: GL and BV; Programmed the web-based CIT: BK and GL; Analyzed the data: BK and GL; Wrote the paper: BV, GL, and BK.

Acknowledgements

Gáspár Lukács received support from the Erasmus+ program for an internship that provided the basis for the current research. Bennett Kleinberg is supported by a grant from the Dutch Ministry of Security and Justice.

References

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

Eom, J.-S., Sohn, S., Park, K., Eum, Y.-J., & Sohn, J.-H. (2016). Effects of varying numbers of probes on RT-based CIT accuracy. *International Journal of Multimedia and Ubiquitous Engineering*, *11*(2), 229–238. http://dx.doi.org/10.14257/ijmue.2016.11.2.23

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 837–845.

Farwell, L. A., & Donchin, E. (1991). The truth will out: Interrogative polygraphy ("lie detection") with event-related brain potentials. *Psychophysiology*, *28*(5), 531–547. http://dx.doi.org/10.1111/j.1469-8986.1991.tb01990.x

Kleinberg, B., & Verschuere, B. (2015). Memory detection 2.0: The First web-based memory detection test. *PLOS ONE*, *10*(4), e0118715. http://dx.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. http://dx.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: Cambridge University Press. Retrieved from http://ebooks.cambridge.org/ref/id/CBO9780511975196A022

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

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

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*, 593–604. http://dx.doi.org/10.1111/psyp.12609

Meijer, E. H., Smulders, F. T. Y., Johnston, J. E., & Merckelbach, H. L. G. J. (2007). Combining skin conductance and forced choice in the detection of concealed information. *Psychophysiology*, *44*(5), 814–822. http://dx.doi.org/10.1111/j.1469-8986.2007.00543.x

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

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

Noordraven, E., & Verschuere, B. (2013). Predicting the sensitivity of the reaction time-based concealed information test: Detecting deception with the concealed information test. *Applied Cognitive Psychology*, *27*(3), 328–335. http://dx.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.

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *7*(6), 657–660. http://dx.doi.org/10.1177/1745691612462588

Peer, E., Samat, S., Brandimarte, L., & Acquisti, A. (2015). Beyond the Turk: An empirical comparison of alternative platforms for online behavioral research. *SSRN Electronic Journal*, http://dx.doi.org/10.2139/ssrn.2594183

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).

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), 1.

Rosenfeld, J. P., Shue, E., & Singer, E. (2007). Single versus multiple probe blocks of P300-based concealed information tests for self-referring versus incidentally obtained information. *Biological Psychology*, *74*(3), 396–404. http://dx.doi.org/10.1016/j.biopsycho.2006.10.002

Rosenfeld, J. P., Soskins, M., Bosh, G., & Ryan, A. (2004). Simple effective countermeasures to P300-based tests of detection of concealed information. *Psychophysiology*, *41*, 205–219.

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. http://dx.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. http://dx.doi.org/10.1111/j.2044-8295.1910.tb00206.x

Suchotzki, K., Verschuere, B., Crombez, G., & De Houwer, J. (2013). Reaction time measures in deception research: Comparing the effects of irrelevant and relevant stimulus–response compatibility. *Acta Psychologica*, *144*(2), 224–231. http://dx.doi.org/10.1016/j.actpsy.2013.06.014

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*, 428–453. http://dx.doi.org/10.1037/bul0000087

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

Please cite this article in press as: Lukács, G., et al. Familiarity-Related Fillers Improve the Validity of Reaction Time-Based Memory Detection. *Journal of Applied Research in Memory and Cognition* (2017), http://dx.doi.org/10.1016/j.jarmac.2017.01.013

Verschuere, B., Ben-Shakhar, G., & Meijer, E. (2011). *Memory detection: Theory and application of the concealed information test*. Cambridge: Cambridge University Press.

Verschuere, B., & De Houwer, J. (2011). Detecting deception within less than a second: Response latency based measures. In B. Verschuere, G. Ben-Shakhar, & E. H. Meijer (Eds.), *Memory detection: Theory and application of the concealed information test*. Cambridge University Press.

Verschuere, B., & Kleinberg, B. (2015). ID-Check: Online concealed information test reveals true identity. *Journal of Forensic Sciences*, http://dx.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. http://dx.doi.org/10.1016/j.jarmac.2015.01.001

Received 10 December 2016;
received in revised form 17 January 2017;
accepted 17 January 2017
Available online xxx

Please cite this article in press as: Lukács, G., et al. Familiarity-Related Fillers Improve the Validity of Reaction Time-Based Memory Detection. *Journal of Applied Research in Memory and Cognition* (2017), http://dx.doi.org/10.1016/j.jarmac.2017.01.013