Acta Psychologica 194 (2019) 7–16

Contents lists available at ScienceDirect

Acta Psychologica

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

Methodological improvements of the association-based concealed information test

Gáspár Lukács*, Ulrich Ansorge

Department of Basic Psychological Research and Research Methods, University of Vienna, Liebiggasse 5, A-1010 Vienna, Austria

ARTICLE INFO

Keywords:
Memory detection
Deception
Concealed information test
Reaction time
Association
ID-EAST

ABSTRACT

The Association-Based Concealed Information Test (A-CIT) is a deception-detection method, in which participants categorize personally relevant items (e.g., their own surnames) as probes together with categorically similar but irrelevant items (e.g., others' surnames) by one key press *A*, while categorizing self-referring "inducer" items (e.g., "MINE" or "MY NAME") with an alternative key press *B*, thereby establishing an association between self-relatedness and *B* and an incongruence between the self-relatedness of probes and *A* (Lukács, Gula, Szegedi-Hallgató, & Csifcsák, 2017). The A-CIT's sensitivity to concealed information is reflected in an incongruence effect: slower responses to probes than to other surnames. To increase the relevance of categories, between trials of the original A-CIT, category-to-response mappings switched or repeated unpredictably. This, however, could have diminished incongruence effects, as the response labels were presented in the corners of the display, veering spatial attention away from the items at screen center. In the present online study ($n = 294$), we therefore tested two improved versions of the A-CIT that do not require spatial attention shifts to and from peripheral labels. One improved version presents per trial only one category label at screen center and requires comparison to the currently presented item. The other improved version is based on the Identification Extrinsic Affective Simon Task (ID-EAST), in which item categorization switches (or repeats) based on colors versus meanings of the central items. Both new versions outperformed the original A-CIT.

1. Introduction

Reliable and valid deception-detection methods are desperately needed, for example, in criminal proceedings and for issues of public security, because without such aids, it is extremely difficult – if not impossible – to tell whether a (potential) perpetrator is telling the truth or lying (Bond & DePaulo, 2006, 2008; Hartwig & Bond, 2011; Kraut, 1980). In the current study, we conceptually replicated a recently introduced deception detection method, the Association-based Concealed Information Test (A-CIT; Lukács, Gula, et al., 2017), and we also introduced two methodically improved alternatives.

In the experiments of the original A-CIT study (Lukács, Gula, et al., 2017), the authors simulated a situation where the authorities suspect the true forename of a person, but this person wants to deny and hide the fact that this forename is hers. For this purpose, participants were tested using either their true forenames, or randomly selected other forenames as "suspected true forenames." The A-CIT task was then used to reveal whether or not the forename in the test is the actual forename of the given participant. This task included two categories of items. The first category consisted of five forenames: the suspected forename as *probe* item, and four other names as *irrelevant* items. The second category consisted of self-referring *inducers*, that is, expressions such as, "mine," "own," or "myself." Items from both categories were presented intermixed in a two-alternative forced choice task: All forenames had to be categorized as "other name" by one key press *A*, while all self-referring expressions had to be categorized as "my name" by an alternative key press *B*. This created a link between self-relatedness and response *B* and, as the name of the participant is self-related, an incongruence between the probe and response *A*: The factually correct semantic category or long-term memory association for an irrelevant name is "other name," while the factually correct category or semantic long-term memory association for the person's own name is "my name." Due to the incongruence between long-term memory associations and task requirements with probes, we expected the "guilty person" (here: each participant in the guilty condition, with his/her own name as a probe) to have larger response conflict when categorizing her own name as "other name," resulting in slower responses and lower accuracy (i.e., rate of correct responses) as compared to the irrelevant other names. At the same time an "innocent person" (here: one presented with another person's name as a probe in a control condition) would

* Corresponding author at: Faculty of Psychology, University of Vienna, Liebiggasse 5, A-1010 Vienna, Austria.
E-mail address: gaspar.lukacs@univie.ac.at (G. Lukács).

https://doi.org/10.1016/j.actpsy.2019.01.010
Received 3 July 2018; Received in revised form 17 January 2019; Accepted 18 January 2019
Available online 25 January 2019
0001-6918/

experience no such conflict, hence not show slower responses or lower accuracy – as was demonstrated in the former study.

However, the authors were concerned that when participants always had to press the same key for the same category, the categorization could become spatial rather than semantic: Examinees would simply recognize the given names as having to be categorized to one side (e.g., always with the key on the left), without forgoing semantic categorization (i.e., regardless of whether the name was their own or not), that is, disregarding the influence of the inducer items.[1] To ensure that the meaning of the categories was more relevant, from trial to trial the "my name" and "other name" category labels switched or repeated positions. This was inspired[2] by the Recoding-Free Implicit Association Test, IAT-RF (Meissner & Rothermund, 2013; Rothermund, Teige-Mocigemba, Gast, & Wentura, 2009) – see Figs. 1 and 2.

1.1. Conceptual replication to assess generalizability

The original A-CIT tested only personal names with corresponding categories that referred to "other name" and "my name." Therefore it is yet to be shown whether the method can be generalized to other items as well (Lukács, Gula, et al., 2017, p. 10). One particular aspect is item saliency: In other RT based CITs, it has been repeatedly shown that probes that have higher personal relevance (e.g., country of origin) can be more easily detected (i.e., the probe-irrelevant differences are larger) than probes with lower personal relevance (e.g., favorite animal; Kleinberg & Verschuere, 2015; Lukács, Kleinberg, & Verschuere, 2017; Verschuere, Kleinberg, & Theocharidou, 2015). Personal names are arguably one of the most relevant probes that could possibly be used, and therefore the validity of the task in case of less salient items is yet to be assessed. Consequently, the first aim of this study was to replicate the validity of the A-CIT in case of country of origin as a high salient probe, and favorite animal as a probe of lower salience. In addition, we also included birthday dates (e.g., "April 19"), which are highly salient, but also contain numbers that may make them more easily distinguishable from the inducer items, and hence could decrease validity if the participant responds based on this visual difference (i.e., whether or not an item contains numbers) instead of based purely on meaning.

This replication was implemented in online settings (Kleinberg & Verschuere, 2015; Peer, Samat, Brandimarte, & Acquisti, 2015), where we could obtain a highly diverse international sample that provides a broad demonstration of generalizability and also more closely reflects the test results of possible criminal suspects than a study involving only university students, although admittedly at the cost of less control and probably generally weaker resulting effects. Furthermore, this allows clearer comparison to other RT-CIT (and IAT) methods that were implemented in the same manner (including the same or similar probe categories; Kleinberg & Verschuere, 2015, 2016; Lukács & Ansorge, 2018; Verschuere et al., 2015, Verschuere & Kleinberg, 2017).

1.2. Introducing two alternatives

The A-CIT in the original study was relatively efficient in discriminating guilty participants from innocent ones, but considering the above-described high-salient probes and the carefully controlled laboratory settings, we expected that it may not compare that well to other recent methods (in particular, see Lukács, Kleinberg, & Verschuere, 2017). However, several possibilities for improvements (e.g., of even online methods) have been noted (Lukács, Gula, et al., 2017, p. 10). For one, it was pointed out that the design might have veered spatial attention away from the crucial (probe and irrelevant) items at screen center due to the labels that were presented in the corners of the display and continually switched places and had to be attended to. Such influences could have substantially increased theoretically uninteresting statistical noise in the data. Therefore, replacing this peripheral label-switching by an alternative design that would still ensure attention to the meaning of the categories, but at the same time does not distract spatial attention from the main items, could boost the incongruence effect.

Here we introduce two separate potential alternatives that may achieve this aim. The first is a simplified version of the original IAT-RF-inspired task. The difference is that the relevance of the semantic categories is ensured by a single category label at the screen center, which precedes each probe or irrelevant item, and which may or may not change from trial to trial (similarly to the switching of the peripheral labels). Participants have to categorize items based on whether or not they belong to this preceding category label (*yes* or *no* response). Hence the *response* labels (*yes* and *no*) are separated from the *category* labels (*mine* or *other*), and remain the same throughout the task, and require no (or very little) attention. This will be referred to as the *Central-label* version, while the original will be referred to as the *Side-label* version.

The second alternative is suggested by the Identification Extrinsic Affective Simon Task (ID-EAST, De Houwer & De Bruycker, 2007), in which, instead of labels, the categorization of the probe and irrelevant items is based on the colors in which they are presented (and which, again, may or may not change from one trial to the next). Hence, no attention to any labels is required, the response choice has to be made based on the presented item alone. The inducer items are presented in the same manner (in varying colors), but they have to be categorized based on meaning only. Consequently, since the inducer, irrelevant, and probe stimuli are randomly intermixed, even if eventually the item is categorized based on color, it first has to be read in order for the participant to decide that the categorization will be based on color (De Houwer & De Bruycker, 2007). In this way, the implicit difference (familiar or not, based on self-relatedness) and the explicit task requirement (response based on the color) is unified within each single item displayed at once. This version will be referred to as *Color-based*.

Note that the different approaches notwithstanding, all three designs have the same basic idea: The explicit instruction (based on peripheral or central label, or color) requires a response to the probe (categorization as "unfamiliar") that is incongruent with the factually correct response in case of a guilty participant ("familiar" category, since the probe is in fact familiar to such a participant).

2. Methods

2.1. Participants

This experiment was run on Figure Eight (www.figure-eight.com; formerly known as CrowdFlower), an online crowdsourcing platform where participants from anywhere in the world can register to complete small online tasks (see Peer et al., 2015). Hence, this website may also be used to offer participation in online experiments by providing a link to the task to be completed (Kleinberg & Verschuere, 2015). People registered on this site as "contributors" complete many such tasks, and their performance may be rated after the completion of the tasks by the "customers" who offered those tasks. Based on these ratings, contributors are categorized into three levels, where contributors with best ratings are categorized as "Level 3." When creating a new task, a customer (in this case, the current authors) may choose the lowest level of contributors that may take the task. We set this to "Level 3", hence, only such "Level 3" contributors were allowed to participate in the study. We opened slots for 300 participants for our experiment, paying 1.30 USD per completed task. After completing the task, a completion password

[1] During pretests, a basic version of the task without switching category-to-response mappings was also piloted but proved rather ineffective.

[2] While the trial structure is similar to that of the IAT-RF, the purpose of this procedural detail in this case is not the same as that of the IAT-RF, which aims to prevent recoding processes (i.e., a simplification of the IAT's double-categorization task to a simple binary classification – as aptly noted by an anonymous reviewer of this manuscript).

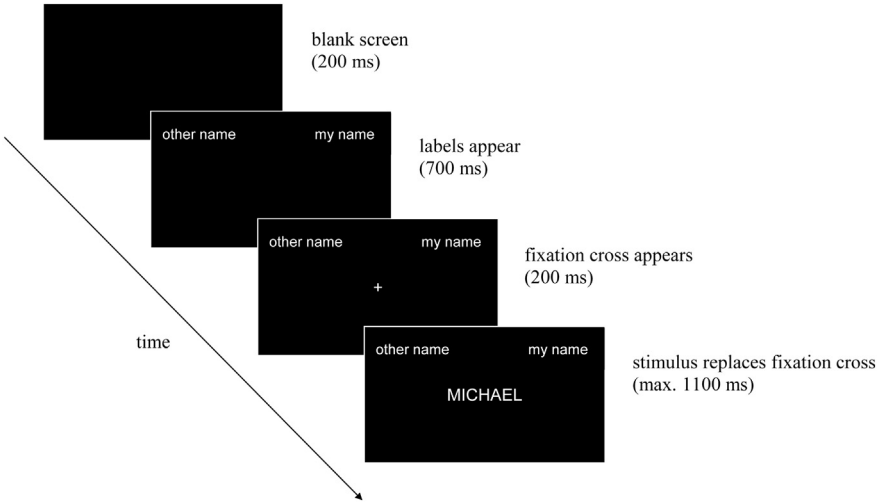

Fig. 1. Example of a trial in the original Association-based Concealed Information Test (A-CIT). First the labels appear, and then follow the stimuli. The next trial begins again with a blank screen, and the subsequent labels either appear at the same locations as on the previous trial or they switch positions.

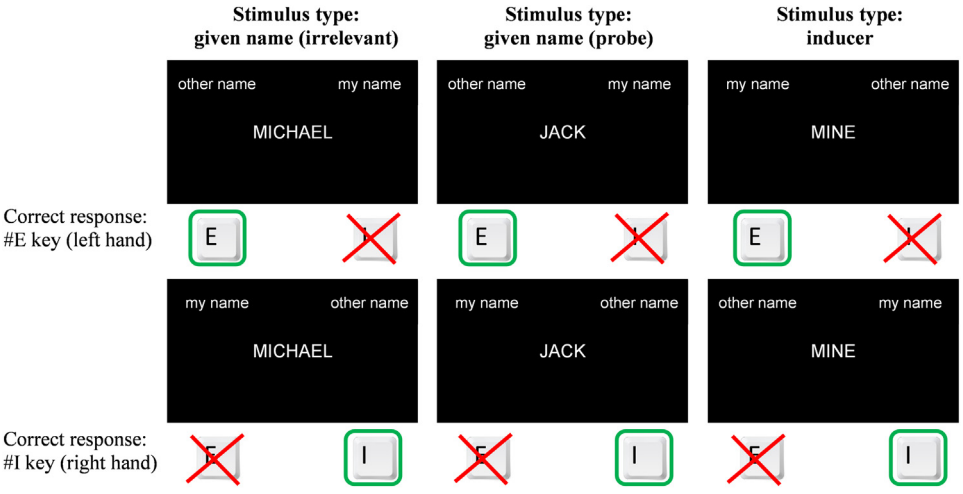

Fig. 2. Examples of the possible stimulus types (own name, here: "Jack" vs. other name, here: "Michael," and probe-related inducer, here: "mine"), label position variations (upper vs. lower row), and corresponding required response keys (keys framed green = currently required response; keys crossed out = currently incorrect or not required response), in the Association-based Concealed Information Test (A-CIT) for a participant called "Jack." Since any of these variations may come up on each trial, the participant has to constantly pay close attention to both the labels and the subsequently presented stimuli to select the correct response key. Note that the presentation and the required response for the probe is exactly the same as for any of the irrelevant other name items. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

was automatically displayed for each participant, which they could submit on the Figure Eight website to receive the payment. Possibly due to simultaneous starting times, 313 participants completed the task.

Previous such online tasks turned out to involve a large amount of invalid participations (e.g., 20–30% in Lukács, Kleinberg, & Verschuere, 2017): Against the instructions, many participants have taken part in the experiment more than once, while a few others have obtained the completion password by cheating, that is, without completing the full task. (We cannot know the exact method of this cheating, we only see that parts of the experimental data are missing from the data of the given participant; in these cases, typically only some of the practice phases were completed. The easiest way however was to manually change the URL to get to the final page. Those more familiar with the HTML/JavaScript framework may also be able to obtain the password by hacking.) Therefore, we have implemented several preventive restrictions. Most importantly, the present task could only be completed in one uninterrupted time from one IP address: Another attempt from an IP address that was already stored with a completed task resulted in a warning prompt on the first page of the task that did not allow continuation. The exclusions due to duplicate IPs accounted for the largest part of the exclusions in previous studies, but thanks to this restriction in our study, there were no duplicates at all, and, hence, no need for any related exclusions. Furthermore, the task was implemented on a single page, preventing manipulation by changing URL to skip parts of the task. The completion password could not be obtained by inspecting the front-end elements of the site either: At the end of the task, the completeness of the data was checked on the server-side and the completion password was returned only in case of positive evaluation. Thanks to these improvements, out of the 313 participants who completed the study, overall only six (1.9%) had to be excluded due to incomplete data.

Each participant was randomly assigned to perform one of the three A-CIT versions: Side-label, Central-label, or Color-based. Each participant was also randomly assigned to the *innocent* or *guilty* condition. In the guilty condition, the probe items were participants' self-reported autobiographical identity details (e.g., their country of origin), simulating a guilty suspect. In the innocent condition, the probe items were not the identity details of the participants (simulating an innocent suspect). Note that the innocent conditions were not strictly needed to test our hypotheses, but are of use to calculate individual detection efficiency receiver operating characteristics (ROCs, see below). The innocents, thus, also serve as controls. They show that arbitrary assignments of specific words to the categories of probes versus irrelevants would not create interference (i.e., a probe-irrelevant performance difference) unrelated to the guilty participants' knowledge that the probes were more self-related than the irrelevants.

Following recent online CIT experiments (e.g., Lukács, Kleinberg, & Verschuere, 2017), we excluded participants with accuracy below 50%

for any of the item types: probes, irrelevants, inducers.[3] In case of the Color-based version, for each of these three item types, performance per each of the two categories was calculated separately (see in Table 2 in the Results section). This resulted in the exclusion of 13 participants (two from guilty Central-label version, eight from guilty Color-based version, three from innocent Color-based version). This left 294 participants; 50 in Side-label version guilty ($M_{age} \pm SD_{age} = 33.18 \pm 8.31$; 72.00% male), 48 in Side-label version innocent ($M_{age} \pm SD_{age} = 36.02 \pm 9.81$ years; 75.00% male), 51 in Central-label version guilty ($M_{age} \pm SD_{age} = 34.33 \pm 10.33$; 74.51% male), 53 in Central-label version innocent ($M_{age} \pm SD_{age} = 34.42 \pm 9.85$; 62.26% male), 45 in Color-based version guilty ($M_{age} \pm SD_{age} = 35.56 \pm 10.22$; 66.67% male), and 47 in Color-based version innocent ($M_{age} \pm SD_{age} = 38.06 \pm 10.86$; 61.70% male).

2.2. Procedure

The entire online application for the three tasks in its original form (except for the removal of the server connections through PHP), as well as a simplified version for demonstration purposes is available online via https://osf.io/ayu4d/.

Upon accessing the link, participants agreed to the informed consent in order to proceed further. (The information included the rule, in boldface font, that at least an upper-intermediate English knowledge is required.) Participants then provided demographic information, and chose, from a dropdown menu, the three autobiographical details that were subsequently used as probes in the A-CIT task: country of origin, date of birth (month and day), favorite animal. This was followed by the very short (three minute) LexTALE English competency test (Lemhöfer & Broersma, 2012), in which 60 words are presented, among which 40 are real English words, while 20 are nonwords, and the instruction is to decide, for each word, whether it is an actual English word or not. This test was implemented as described at www.lextale.com, with the only difference that a 4-s time limit applied to each response to curb possible cheating (i.e., looking up the words online or in a dictionary during the task). The LexTALE minimum score for upper intermediate (B2) level is 60% accuracy (Lemhöfer & Broersma, 2012, p. 341). Consequently, those who did not achieve a score above our more lenient threshold of 56% clearly did not have the required English skill, and therefore were automatically disqualified and redirected to the Figure Eight website. This screening was important due to the tasks' reliance on semantic associations, which requires a clear understanding of basic English.

Then followed one of the A-CITs as described below. After the task, there was a short survey where participants rated the personal importance of the items used in the task (their country of origin, birthday, and favorite animal; on a scale from one to six, where one is "entirely unimportant" and six is "very important"), and finally the participants were given a brief explanation about the purpose of the study. The entire experiment took 20–30 min, within which the A-CIT took around 15 min.

2.3. The association-based concealed information tests

Participants were informed that the following task simulates a lie detection scenario, during which they should try to hide their identities. They were also told that they may actually not see their own details in the task, in which case they are in the "innocent" condition, simulating an innocent suspect. They were then presented a short list of items within each of the three categories in the task (countries, dates, animals). The items on this list never contained any of the actual identity details of a given participant. The participants were asked to choose any (but a maximum of two per category) items that were personally meaningful to them or in any way appeared different from the rest of the items on those lists. Subsequently, the 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 innocent condition, five items were selected for each of the three categories. Out of these five, one was randomly assigned to be a probe, while the remaining four served as irrelevant items. This assignment was not known to the participant, but served only to have the same arrangement of item types as in the guilty condition, for the subsequent statistical analysis. For a participant in the guilty condition, their self-reported identity details served as the probe item in each of the four categories, while a random four of the non-chosen items served as irrelevant items. Thus, in either condition, there were altogether 15 unique items: three probes, and 12 irrelevant items.

2.3.1. Side-label version

The Side-label version was a replication of the original study (as described in the Introduction), with only some minor differences: (1) The labels and inducers were not referring to personal names, but to generic self-relation ("mine" vs. "other"; see Appendix), (2) the labels here were displayed at the bottom instead of the top, (3) the background color was not fully black but very dark blue (almost black; RGB values: 3, 17, 22), (4) the interstimulus-interval varied randomly between 200 and 300 ms instead of a fixed 200 ms, (5) there was a "False!" feedback in case of an incorrect response (while in the original A-CIT there was no such feedback; either the correct response or the end of the response window was awaited – in the latter case "Too slow!" was displayed, same as in our replication), and (6) each block included 136 regular trials instead of 137 which (in the original A-CIT) included an extra first trial that was a randomly chosen inducer (omitted from all analyses).

2.3.2. Central-label version

The Central-label version is very similar to the Side-label version, with the only difference that the response labels are the same throughout each trial, and the relevance of the semantic categories is ensured by a category label at the screen center (instead of two on the two sides) that precedes each main item. From trial to trial, this single semantic category label displayed in the middle of the screen switches or repeats between "other" and "mine" categories. As in the Side-label version, a probe, an irrelevant, or an inducer item follows. (These items are identical to those in the Side-label version, see Appendix.) When this item appears, the participant has to decide whether or not this item belongs to the semantic category of the label, pushing one key for *No* and another key for *Yes*. According to the instructions, the probe and the irrelevant items all belong to the "other" category, while inducers referring to the probe belong to the "mine" category. For example, in case of testing for the processing of a person's true name (as probe), if the first expression "mine" is followed by the irrelevant item "MICHAEL" or probe item "JACK," then in either case the correct response is *No*, but if it is followed by the probe-referring inducer "MY NAME," then the correct response is *Yes*. However, if the first

[3] We have also performed the analyses using two alternative exclusion criteria, a more liberal one, and a more conservative one: (1) no exclusions at all (i.e., including all 307 participants who completed the test), and (2) excluding Tukey's outliers as in the original A-CIT study (Lukács, Gula, et al., 2017; accuracy rate over 1.5 interquartile outside the interquartile range for either the main items [probe or irrelevant] or for the inducer items, in which case here the lowest accuracy for main items was 67.5%, while the lowest accuracy for inducers was 62.5%; 285 participants remained). There is no relevant change in any of our findings when using either of these two alternative methods, except that, when using Tukey's outliers (less data included), the RT median difference between the Side-label and Central-label versions loses statistical significance, presumably due to insufficient statistical power for this relatively small effect size; $t(96) = 1.74$, $p = .086$, $d_{between} = 0.35$. The results per subject for both these alternative exclusions methods are available via https://osf.io/ayu4d/ (along with the one used for the reporting in this manuscript).

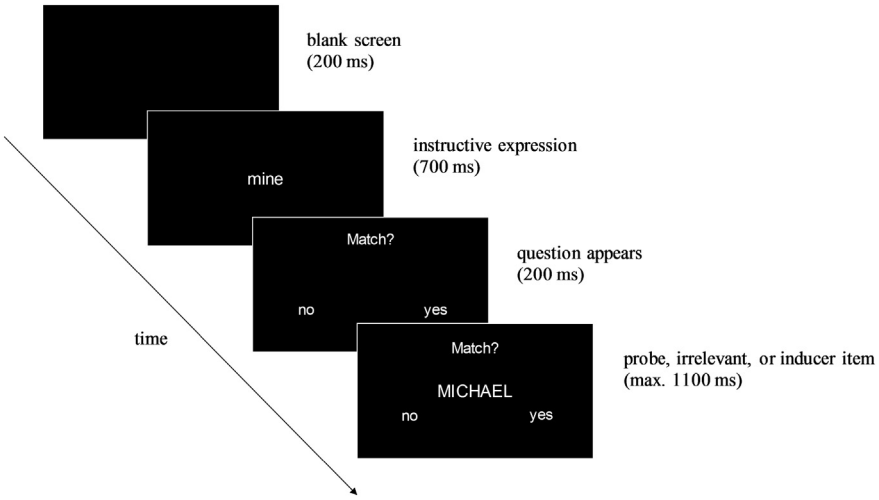

Fig. 3. Example of a trial in the Central-label version. First, an instructive expression (or category label) is displayed (here: "mine"), and afterwards the item that requires the response (here: "MICHAEL"; cf. Fig. 1, and perhaps see also Sriram & Greenwald, 2009). (Note that the "Match", "no", and "yes" texts never change positions; they serve only as reminders and thus should draw much less attention than the peripheral labels in the original version of the task.)

Table 1
Required responses in the central-label version.

Instructive expression	Probe-related (*my name*)			Irrelevant-related (*other name*)		
Following item	Irrelevant (*MICHAEL*)	Probe (*JACK*)	Inducer (*MINE*)	Irrelevant (*MICHAEL*)	Probe (*JACK*)	Inducer (*MINE*)
Required response	No (*#E* key)	No (*#E* key)	Yes (*#I* key)	Yes (*#I* key)	Yes (*#I* key)	No (*#E* key)

Note. Stimulus types ("Following item") and required responses depending on the preceding instructive expression (or category label), with examples in parentheses. Note that the participant's long-term memory of the probe (i.e., his/her name) suggests a categorization and response that is always incongruent to the required categorization and response.

expression is "other," then the correct response for either "MICHAEL" or "JACK" is *Yes*, and the correct response for "MY NAME" is *No* (see Fig. 3 and Table 1). Here the hypothesis was that guilty participants would always have slower responses and lower accuracy for the probe compared to the irrelevant items, since the required response to the probe is always incongruent with the facts: If the probe is the actual name of the given examinee, "mine" followed by the probe would suggest to the participant to respond *Yes*, but the task requires the response *No*, and, similarly, "other" followed by the probe would suggest to the participant to respond *No*, but the task requires *Yes*.

2.3.3. Color-based version

In the Color-based version, the relevance of response category labels is again almost entirely eliminated, serving only as reminders: a "not mine"[4] label always displayed on the left side, "mine" on the right side. Instead of switching or changing labels, the colors of the stimuli change randomly, but a semantic processing of all items is still required, as in case of the inducers, the stimuli have to be categorized based on meaning (in our study, the standard computer keyboard #E key on the left for "not mine," and #I key on the right for "mine"), while in case of the probes and irrelevants only, they have to be categorized based on their color (e.g., #E key for red, and #I key for green). Since the stimuli are randomly intermixed, even if eventually the probe is categorized based on color, the examinee first has to read it and consider its meaning in order to decide that the categorization will be based on color. Since inducers have to be categorized here as self-referring ("mine") or other-referring ("not mine"), this version included not only self-referring, but also other-referring inducers (see Appendix for details). Notably, due to this double-categorization task, this A-CIT version, compared to the other two, is much closer in both design and in concept to the IAT that inspired it (De Houwer & De Bruycker, 2007).

The hypothesis here is that categorization of guilty participants is slower and also more often incorrect for probes compared to the irrelevant items, as the probes are semantically more similar to the self-related category and, thus, delay the decision about which task to perform, that is, that color has to be processed and responded to.

2.3.4. Summary and implementation of the tasks

Note that all three versions had an *instruction component* based on which the items had to be categorized (through pressing either #E on the left or #I on the right). In the Side-label version, category-response mappings were flexible. For example, participants had to press #I on the right when the "mine" label appeared on the right (with the "other" label on the left), and the main item was a self-referring item (e.g., "my country"). In the Central-label version, category-response mappings were fixed. For example, participants always had to press #I for "yes" on the right when the main item was preceded by the singular centrally presented "mine" label, and the main item was a self-referring item, as pressing the #I key corresponded to the response "yes", confirming that the label and the following item belong to the same category. In the Color-based version, category-response mappings were also fixed. For example, participants had to press #I on the right when the main item was a self-referring item (regardless of color), or when it was a probe or irrelevant item in green color (and if the probe or irrelevant item was red, they had to press the #E on the left).

In the main task of each of the three versions, each trial began with a blank screen for 200–300 ms (random length between a minimum of 200 ms and a maximum of 300 ms). After this, in case of the Side-label version, both labels appeared on the lower part of the screen. After another 700 ms, a fixation cross appeared in the middle of the screen for another 200–300 ms, in order to draw the participant's attention to the oncoming item, which appeared afterwards. In case of the Central-label version, one label appeared in the middle of the screen. After another 700 ms, the label disappeared, and the question "Match?" appeared for 200–300 ms at the top of the screen to prepare the participant for responding to the oncoming stimulus. (As reminders of the corresponding

[4] We decided by "not mine" instead of "other" in this case, because, lacking any alternative word that clearly expresses this relation, all other-referring inducer expressions included the word "other." By giving an alternative wording in the response label as "not mine," we tried to limit the focus on the single word "other," and increase attention to the actual concept of "other" as the opposite of "mine", i.e., "not mine." Such consideration was not required in the other two versions, since they did not involve other-referring inducers, and therefore the simplest straightforward label for the "other" category could just be the very word "other."

keys, a “no” on the left and a “yes” on the right were always displayed at the bottom part of the screen alongside with the question, but it was not strictly necessary to attend to the labels, as these remained the same throughout the experiment.) Then again, the main item appeared in the middle. In case of the Color-based version, there was no label; the items simply appeared one after the other with 200–300 ms intervals in-between. As reminders of the corresponding keys, a “not mine” in red on the left and a “mine” in green on the right were always displayed at the bottom part of the screen. Again these labels did not change positions throughout the experiment requiring little spatial attention if at all. In each version, the participant had 1100 ms to respond to the item (Lukács, Gula, et al., 2017, p. 5). In case of an incorrect or too slow response, corresponding feedback was displayed for 400 ms (captions: “False!” and “Too slow!”).

The main task was preceded by a comprehension check and two practice blocks. The check served to ensure that the participant had understood the task. The items consisted of 10 trials with 10 different main items (each of which was randomly chosen from one out of the three categories), including two probe and eight irrelevant items, with random instruction component. In this task, participants had plenty of time (10 s) to choose a response – however, each trial required a correct response. In case of an incorrect response, the participant immediately got a corresponding feedback, was reminded of the instructions, and had to repeat the trial. This check guaranteed that the eventual differences (if any) between the responses to the probe item and the responses to the irrelevant items were not due to misunderstanding of the instructions or any uncertainty about the required responses in the eventual task.

In the following first practice block, the response window was longer than in the main task (2 s instead of 1100 ms), while the second practice block had the same design as the main task. Both practice blocks consisted of 16 trials. In either practice block, in case of too few valid responses, the participants received a corresponding feedback, were reminded of the instructions, and had to repeat the practice block. The constraints were: (1) more than a quarter of responses have a reasonable reaction time (above 200 ms; to ensure the participant is not pressing keys randomly), (2) at least half of the responses is valid (neither incorrect nor too slow, i.e., over the limit of the response window of 1100 ms or 2 s) for each of the two response keys (to ensure that the participant is not focusing on either side), and (3) at least two thirds of the responses for inducers, and two thirds of the responses for main items are valid.

The following main task consisted of three blocks of 136 trials each, one block for each category (countries, dates, animals; in random order), including 80 trials with actual details (each of the five details in each category 16 times) and 56 with inducers (14 times the four self-referring expressions, or, in case of Color-based version, seven times two self-referring and seven times two other-referring expressions); thus, altogether 408 trials in the main task. The inducers always corresponded to the category in the given block. For example, for dates, they may have been “mine” or “my birthday” for self-referring items or, in case of Color-based version, an other-referring item could be “other date.” All stimuli were presented in random order but with several restrictions (to avoid word repetition and to balance the instruction component types and stimulus categories).[5] There were breaks between the blocks during which participants could take a rest and continue when they felt ready.

2.4. Data analysis

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

Along with the conventional values reported for *t*-tests, we also report Cohen's *d* values following the formula given in recent RT-CIT studies (Kleinberg & Verschuere, 2015, 2016; Lukács, Kleinberg, & Verschuere, 2017; Verschuere et al., 2015; adopted from Lakens, 2013).

To assess the efficiency of discriminating between guilty and innocent conditions, we calculated areas under the ROC curve (AUC – area under the curve; a diagnostic efficiency measure, for binary classification, that takes into account the distribution of all predictor values; e.g., Zou, O'Malley, & Mauri, 2007). The AUC can range from 0 to 1, where 0.5 means chance level classification, and 1 means flawless classification (i.e., all guilty and innocent classifications can be correctly made based on the given predictor variable, at a given cutoff point). RT-CIT studies usually use mean RTs and accuracies as the basis of predictor variables: More precisely, the difference between the mean RT to probes and the mean RT to irrelevant items, and the difference between the accuracy rate to probes and accuracy rate to irrelevant items, calculated for each individual (Seymour, Seifert, Shafto, & Mosmann, 2000; Verschuere, Crombez, Degrootte, & Rosseel, 2010). Here, we use median RTs as RT-based predictor, since it was found to be superior to mean RTs in the first A-CIT study (Lukács, Gula, et al., 2017) as it was expected due to lower sensitivity to outliers and skewness (e.g., Ratcliff, 1993, pp. 522, 531).

For the Color-based version, we expected guilty participants to have slower responses and lower accuracy to the probe compared to the irrelevant items when having to categorize the probe to the side that matches the key for the other-referring category of the inducers, but faster responses and higher accuracy when having to categorize it to the side that matches the key for the self-referring category of the inducers. Correspondingly, the main predictors for RT medians and accuracy rates were calculated as the sum of the following two compounds: (1) the conventionally calculated probe-irrelevant difference (probe values minus irrelevant values) for items categorized opposite to self-referring inducers (i.e., together with other-referring inducers) and (2) a reverse probe-irrelevant difference (irrelevant values minus probe values) for items categorized together with self-referring inducers. Additional separate analyses of probe-irrelevant differences are reported for both of these categorization types.

We used an alpha level of 0.05 for all statistical significance tests.

3. Results

The results data for the experiment can be retrieved from the Open Science Framework data repository via https://osf.io/ayu4d/ (Open Science Collaboration, 2012).

3.1. Manipulation check

The ratings of personal importance showed the expected differences (Kleinberg & Verschuere, 2015): Both countries of origin (rating = 5.19 ± 1.05) and birthdays (rating = 4.99 ± 1.29) were reported as more personally relevant than favorite animals (rating = 4.72 ± 1.23), $t(263) = 5.52$, $p < .001$, $d_{between} = 0.34$; $t(263) = 2.88$, $p = .004$, $d_{between} = 0.18$. In addition, ratings for countries were significantly higher than for birthdays, $t(263) = 3.13$, $p = .002$, $d_{between} = 0.19$.

[5] The same item was never repeated on consecutive trials. The same type of instruction component was never repeated on more than three consecutive trials. Each main item (the probe, and the four irrelevants) was preceded, in 50% of its appearances, by another main item, and in the other 50% of its appearances, by an inducer. Furthermore (and also within each of the two cases described in the previous sentence), each main item was accompanied by the two possible instruction types equally often (e.g., for Side-label version, 50% one label position, 50% the other). The inducers were, on average, also accompanied by the two possible instruction types on equal numbers of trials.

Table 2
RT medians, accuracy rates, Cohen's Ds, and areas under the curves, in each condition.

	Side-label		Central-label		Color-based	
	Innocent	Guilty	Innocent	Guilty	Innocent	Guilty
Medians (ms)						
Probe-other	743 ± 115	752 ± 97	666 ± 95	689 ± 100	649 ± 72	717 ± 99
Probe-self					631 ± 67	643 ± 81
Irrelevant-other	747 ± 115	734 ± 97	657 ± 82	651 ± 99	655 ± 68	664 ± 73
Irrelevant-self					626 ± 70	645 ± 79
Inducer-self	772 ± 101	773 ± 94	713 ± 74	712 ± 84	658 ± 66	678 ± 71
Inducer-other					707 ± 79	718 ± 69
d_{within}	−0.12	0.48	0.23	0.69	−0.24	0.81
$d_{between}$	0.63		0.61		1.13	
AUC	.668		.667		.781	
Accuracies (%)						
Probe-other	89 ± 08	88 ± 07	85 ± 11	80 ± 11	91 ± 08	79 ± 13
Probe -self					90 ± 08	91 ± 08
Irrelevant-other	89 ± 08	91 ± 05	84 ± 11	85 ± 08	91 ± 07	90 ± 06
Irrelevant-self					91 ± 06	88 ± 08
Inducer-self	89 ± 08	87 ± 09	84 ± 10	84 ± 09	88 ± 08	84 ± 11
Inducer-other					82 ± 12	83 ± 09
d_{within}	−0.05	−0.63	0.13	−0.59	0.23	−1.04
$d_{between}$	−0.62		−0.80		−1.46	
AUC	0.668		0.735		0.848	

Note. Means and *SD*s (in the format of MEAN ± *SD*) for individual median RTs and accuracies (percentages of correct responses) for *Probe-Other* (i.e., "probe items" according to the rest of the text, e.g., participant's own country of origin, where *Probe-Other* denotes probes categorized opposite to self-related expressions) and *Probe-Self* (realized only in case of Color-based version, for those probes categorized together with self-related expressions), *Irrelevant-Other* ("irrelevants" according to the rest of the text; categorized opposite to self-related expressions) and *Irrelevant-Self* (realized only in case of Color-based version, for those irrelevants categorized together with self-related expressions), *Inducer-Self* ("inducers" according to the rest of the text, i.e., self-referring expressions), *Inducer-Other* (only in case of Color-based version; other-referring expressions). Cohen's *d* effect sizes: d_{within} for probe-irrelevant differences, $d_{between}$ for guilty-innocent differences for each A-CIT version. *AUC*: Area under the curve, that is, the efficiency of classifying participants as the guilty or innocent in each A-CIT version.

3.2. Group-level analysis

All means and *SD*s of individual RT medians and response accuracies, for the different stimulus types, in guilty and innocent conditions, are given in Table 2.

We conducted an analysis of variance (ANOVA), with the between-subjects variables Version (Side-label, Central-label, and Color-based) and Guilt (guilty and innocent conditions) on probe-irrelevant differences in RT medians and in accuracy rates. The ANOVA revealed a significant interaction for both RT medians, $F(2, 288) = 5.94$, $p = .003$, $\eta_p^2 = 0.040$, and accuracy rates, $F(2, 288) = 17.48$, $p < .001$, $\eta_p^2 = 0.108$.

We then conducted a one-way ANOVA, for guilty conditions only, with the between-subjects variable Version (Side-label, Central-label, Color-based) on probe-irrelevant median RT differences, which revealed a significant main effect, $F(2, 143) = 5.35$, $p = .006$, $\eta_p^2 = 0.070$. Further *t*-tests show that, compared to the Side-label version, guilty participants have larger probe-irrelevant RT differences both in the Color-based version, $t(93) = 3.25$, $p = .002$, $d_{between} = 0.67$, and in the Central-label version, $t(99) = 2.01$, $p = .047$, $d_{between} = 0.40$. However, there is also no significant difference for this measure between the Central-label version and the Color-based version, $t(94) = 1.42$, $p = .159$, $d_{between} = 0.29$.

To examine the effects of item category and its possible interactions across A-CIT versions, we ran a mixed-design ANOVA, with Item Category (countries, dates, or animals) as within-subject factor and Version (Side-label, Central-label, Color-based) as between-subjects factor. Neither the main effect of Item Category nor the interaction of Item Category × Version proved significant for median RT probe-irrelevant differences, $F(2, 286) = 2.96$, $p = .054$, $\eta_p^2 = 0.020$; $F(4, 286) = 1.52$, $p = .195$, $\eta_p^2 = 0.021$. Paired sample *t*-tests, however, show that the median RT probe-irrelevant differences, overall in the three groups, were larger for the country items than for the animal items – but with a very small effect, $t(145) = 2.17$, $p = .031$, $d_{between} = 0.18$. The dates-animals and countries-dates comparisons show no significant differences between the (significant) RT probe-irrelevant differences, $t(145) = 0.96$, $p = .338$, $d_{between} = 0.08$; $t(145) = 1.40$, $p = .163$, $d_{between} = 0.12$.

Similarly for probe-irrelevant accuracy rate differences, we conducted a one-way ANOVA, for guilty conditions only, with the between-subjects factor Version (Side-label, Central-label, Color-based), which revealed a significant main effect, $F(2, 143) = 18.36$, $p < .001$, $\eta_p^2 = 0.206$. Here, the follow-up *t*-tests show that guilty participants have larger probe- irrelevant accuracy rate differences in the Color-based version than either in the Side-label version, $t(93) = 5.15$, $p < .001$, $d_{between} = 1.06$, or in the Central-label version, $t(94) = 4.39$, $p = .001$, $d_{between} = 0.90$. There is no significant difference for this measure between the Central-label version and the Side-label version, $t(99) = 0.74$, $p = .463$, $d_{between} = 0.15$.

Same as for RTs, the mixed-design ANOVA on accuracies, with Item Category (countries, dates, or animals) as within-subject factor and Version (Side-label, Central-label, Color-based) as between-subjects factor, showed no significant differences for the main effect of Item Category or for the interaction of Item Category × Version, $F(2, 286) = 1.14$, $p = .321$, $\eta_p^2 = 0.008$; $F(4, 286) = 0.76$, $p = .542$, $\eta_p^2 = 0.011$. Paired sample *t*-tests also did not give significant differences between any two of the three item categories ($p > .09$).

We conducted an additional analysis, in case of the Color-based version, for probe and irrelevant items categorized to the same side as other-related expressions and those categorized to the same side as self-related expressions. For items categorized to the same side as other-related expressions, there was a large significant effect both between probe and irrelevant RT medians, $t(44) = 5.83$, $p < .001$, $d_{within} = 0.87$, and between probe and irrelevant accuracy rates, $t(44) = 6.40$, $p < .001$, $d_{within} = 0.95$. However, for items categorized to the same side as self-related expressions, there was no such significant effect between probe and irrelevant RT medians, $t(44) = -0.36$, $p = .720$, $d_{within} = -0.05$. The differences between probe and irrelevant accuracy rates were statistically significant, $t(44) = 2.67$, $p = .011$, though showing a much smaller effect,

$d_{within} = 0.40$, than items categorized to the same side as other-related expressions. Finally, for completeness, we directly compared the probe-irrelevant differences for items categorized to the same side as other-related expressions with the probe-irrelevant differences for items categorized to the same side as self-related expressions. This difference was significant and large for both RTs, $t(44) = 5.41$, $p < .001$, $d_{within} = 0.81$, and for accuracy rates, $t(44) = 7.19$, $p < .001$, $d_{within} = 1.07$.

For innocent participants, as expected, the RT or accuracy differences between probes and irrelevants were never significant ($p > .1$). Nonetheless, for completeness – while not theoretically interesting – we also conducted a one-way ANOVA for innocent conditions only, for probe-irrelevant median RT and accuracy differences, between the three Versions (Side-label, Central-label, Color-based). Here, we did have a positive finding for median RTs: $F(2, 145) = 3.31$, $p = .044$, $\eta_p^2 = 0.044$. Follow-up t-tests revealed only one significant difference; larger RT median probe-irrelevant differences for the Central-label than for the Color-based version: $t(98) = 2.35$, $p = .021$, $d_{between} = 0.47$. This is clearly a false positive finding, given the probe-irrelevant RT (and accuracy rate) differences themselves were not significant, and since probes for innocents are chosen randomly from among irrelevants.[6] There were no significant differences between any other two of the three versions ($p > .7$). There were no significant differences between the probe-irrelevant accuracy rate differences across the versions, $F(2, 145) = 1.32$, $p = .271$, $\eta_p^2 = 0.018$; with $p > .1$ for the t-tests for differences between any two of the three versions.

We report Spearman–Brown odd-even split-half reliability following other RT-CIT studies (see in particular Kleinberg & Verschuere, 2016; Spearman, 1910; Brown, 1910). The reliability coefficient in the guilty condition was $\rho = 0.66$ for the Side-label version, $\rho = 0.63$ for the Center-label version, and $\rho = 0.30$ for the Color-based version. In innocent condition (where the probe was a random item, hence, the probe-irrelevant difference was also random), it was $\rho = 0.22$ for the Side-label version, $\rho = 0.17$ for the Center-label version, and $\rho = -0.24$ for the Color-based version.

3.3. Individual classification

Probe-irrelevant differences in median RTs and accuracies were used as predictor variables to calculate AUCs, which are shown for each condition in Table 2. Both RT and accuracy have medium to large guilty-versus-innocent between group differences in all versions. Therefore, as an aggregate predictor for each version, we also computed a logistic regression with guilty or innocent as the outcome predicted from the probe-irrelevant RT median and accuracy rate differences (again following Lukács, Gula, et al., 2017). Assessment of goodness-of-fit revealed a significant improvement relative to a constant-only model for all versions, Side-label: $\chi^2(2, N = 97) = 11.4$, $p = .003$, Nagelkerke's $R^2 = 0.192$; Central-label: $\chi^2(2, N = 103) = 14.8$, $p = .001$, Nagelkerke's $R^2 = 0.255$; Color-based: $\chi^2(2, N = 91) = 24.0$, $p < .001$, Nagelkerke's $R^2 = 0.587$. In all three A-CIT versions, the probability of guilt was significantly associated with probe-irrelevant median RT differences as individually contributing predictors (Side-label: $B = -16.05$, Wald χ^2 [1] = 5.3, $p = .021$; Central-label: $B = -11.45$, Wald χ^2 [1] = 5.0, $p = .025$; Color-based: $B = -17.91$, Wald χ^2 [1] = 9.9, $p = .002$; median RTs in seconds for coefficient readability). Accuracy rates were also significant contributors in all versions (Side-label: $B = 10.05$, Wald χ^2 [1] = 5.0, $p = .025$; Central-label: $B = 12.77$, Wald χ^2 [1] = 10.0, $p = .002$; Color-based: $B = 12.74$, Wald χ^2 [1] = 16.4, $p < .001$).

The AUC for the model-based predicted probability of "guilt" was 0.721 CI [0.618, 0.824] for Side-label version, 0.758 CI [0.664, 0.851] for Central-label version, and 0.889 CI [0.822, 0.957] for Color-based version.

[6] This difference, unlike any of our other findings, is not significant when using either of the two alternative exclusion criteria (i.e., when including either less or more participants).

4. Discussion

In the present paper, we have introduced two improved versions of the A-CIT, along with conceptually replicating the original one (Lukács, Gula, et al., 2017). The conceptual replication was successful: Although in the current less controlled online settings and less salient probes yielded a lower guilty-innocent classification accuracy than the original study, the results are comparable with those in the standard RT-CIT (i.e., "single-probe" version, where probes are presented sequentially, same as in the A-CIT; Lukács, Kleinberg, & Verschuere, 2017; Verschuere et al., 2015). One new version is still mainly based on the IAT-RF concept (Rothermund, Teige-Mocigemba, Gast, & Wentura, 2009) but in a simplified design. While this version outperformed the original, it still did not achieve a very high classification accuracy. Nonetheless, it could still be theoretically interesting for future studies, and it also opens up new possibilities for further development (see below). The other new version is based on the ID-EAST (De Houwer & De Bruycker, 2007). This version also outperformed the original A-CIT, yielding a classification accuracy (AUC = 0.89) comparable to the most accurate deception-detection methods (Meijer, Verschuere, Gamer, Merckelbach, & Ben-Shakhar, 2016).

Interestingly, we found no substantial effects for the item types with different levels of salience (personal relevance), which was shown to have large effects in case of the standard RT-based CIT (Verschuere et al., 2015). This can be explained by the premise that the A-CIT relies merely on associations induced by the task: Regardless of personal relevance, the context of deception detection, which the examinee is made aware of through inducers, elicits the probe-irrelevant RT and accuracy rate differences. This may be a very relevant advantage in case of real-life application, where there may be less salient, peripheral details (e.g., objects in the room where a crime was committed), and, thus, a method that is equally sensitive to peripheral as to central details has more applicable probe items, and, hence, more cases where it may be used (about the scarcity of applicable probe items in general, see Podlesny, 2003).

Apart from providing valid deception detection methods, our designs may also be of interest to researchers of implicit associations. An IAT-RF simplified in a similar way as in our study may be directly compared to the original IAT-RF to assess whether it works as well as in our study, and if perhaps there is an increase in predictive power for implicit associations. Moreover, unlike the standard IAT, these tasks involve not only two but multiple (and in fact possibly any number of) items within a relatively short task, and our results prove that they can be reliably used to reveal (concealed) associations of these items. This may be useful when the implicit associations of multiple concepts are being examined, as, for example, affinity with any of the several parties of a current election (Bluemke & Friese, 2008).

4.1. Limitations and direction for future research

In case of the Color-based version, one of the findings may be surprising. Our prediction was that guilty participants would have slower responses and lower accuracy to the probe compared to the irrelevant items when having to categorize the probe to the side that matches the key for the other-referring category of the inducers and faster responses and higher accuracy when having to categorize it to the side that matches the key for the self-referring category of the inducers. The slower responses and lower accuracy for the other-referring side categorization were as expected, with a very large effect. However, for the self-referring side, we found no RT difference at all between the probe and the irrelevants. We did find, though, a significant effect for the accuracy in the expected direction (higher accuracy for probes), but

even so the effect size was fairly small compared to the opposite categorization or to the similar accuracy effect sizes in the other two A-CIT versions. A likely explanation for these null and suboptimal findings is the fact that, regardless of associations, the probe is still a rare item standing out among the other, irrelevant items in its category, and hence elicits a response conflict, which in turn impedes the expected faster responses and higher accuracy. One solution for this could be shortening the deadline for responses for this self-referring side, which may encourage the participant to suppress the conflict and make a quick response. Alternatively, the proportion of items to be categorized to this side could be lowered to increase the target nature of the probe inasmuch as the self-related items appear less frequently overall, hence, creating an even larger conflict in case of its appearance, leading to even slower responses in case of the other-referring side categorization (also reflecting the mechanism of the standard RT-based CIT; Suchotzki, Verschuere, Van Bockstaele, Ben-Shakhar, & Crombez, 2017; Varga, Visu-Petra, Miclea, & Buş, 2014).

It is also possible that the lacking facilitation by probes categorized as self-related inducers in the Color-based CIT reflected a floor effect and that participants were already as fast as they could get when responding to the irrelevants. In any case, because in the Color-based CIT we observed the same interference by incongruent probes (i.e., probes categorized together with the other-related inducers) relative to congruent probes (i.e., probes categorized together with the self-related inducers) as between incongruent probes and irrelevants, the latter interference cannot be due to potentially existing word (processing) differences.

To note, same as in almost all CIT studies, a potential limitation in our experimental design is that in probe-irrelevant comparisons of both other currently tested A-CIT versions, probe and irrelevant conditions also differ in terms of the words used, such that word differences could theoretically contribute to probe-irrelevant performance differences. That this happened is, of course, not very likely given that (1) the same words that figured as probes and irrelevants for one participants could have had reversed roles and functioned as irrelevants and probes, respectively, for another participant, and (2) among innocent controls interference by probes relative to irrelevants was absent although these participants got a random selection of the same words as probes versus irrelevants as the guilty participants. However, the interference by incongruent relative to congruent probes among guilty participants in the Color-based A-CIT seems to seal the case in favor of the probe's self-relevance as the factor responsible for the interference, as this interference is observed between conditions using the same words (i.e., both conditions using the probes).

Another potential limitation is the low split-half reliability that we obtained from our data, especially in case of the Color-based version. However, we do not have sufficient numbers of trials to obtain conclusive test-retest reliability coefficients (Brown, 1910; Spearman, 1910). In particular, each probe detail is presented only eight times in the RF-based versions (with altogether three times eight, i.e., 24 probes), and only mere four times in the EAST-Based version (with altogether three times four, i.e., 12 probes) – with 10–20% of trials excluded for incorrect or too slow responses (Table 2). Kleinberg and Verschuere (2016; Fig. 1 and Fig. 2, Exp. 2) show that the split-half reliability coefficients are sharply increased when using the first eight blocks that include 16 repetitions of probe details (altogether 48 probes; average ρ around 0.45), as compared to using the first four blocks that included eight repetitions of each probe detail (analogous to our Side-label and Center-label versions; altogether 24 probes; average ρ around 0.25), and especially compared to using only the first two blocks that included four repetitions of each probe detail (analogous to our Color-based A-CIT; altogether 12 probes; average ρ around −0.19, i.e., implausible negative correlation) – with, however, very little changes in the corresponding AUCs regardless of the number of blocks (and trials) taken into account (averaging around 0.63, 0.64, and 0.68, respectively). By comparison, our reliability coefficients (in the guilty conditions) are in fact very high for all A-CIT versions in consideration of the corresponding probe trial numbers. Still, to better assess test-retest reliability, we would need to present the task at least twice. This may be addressed in future studies. On the other hand, this measure is not strictly necessary for the evaluation of these tests: The general reliability of such tests is attested to by the obtained AUCs. Thus, it is no wonder that in the vast literature of the CIT extremely few articles report separate reliability coefficients (and these coefficients vary wildly from ρ = 0.15 to ρ = 0.79 depending on the experimental design and settings: Noordraven & Verschuere, 2013, Kleinberg & Verschuere, 2015, 2016, Lukács, Kleinberg, & Verschuere, 2017).

Both novel A-CIT versions also leave many possibilities for improvements that could increase their guilty-innocent classification accuracy even further. For example, in case of the Central-label version, an additional enhancement of the involved categorization processes could be achieved by a secondary task: categorizing each label (appearing in the middle of the screen in the beginning of the task; see Fig. 3) overtly as belonging to the other-related or self-related categories (“other” or “mine”) by pressing keys assigned to the categories: for example, a key on the left for “other” and a key on the right for “mine.” In the given example of this case, whatever response the upcoming main item requires, a *Yes* or a *No*, a guilty person will then less easily categorize (his or her own name) with the key on the left. As was the case with the current diminution of theoretically uninteresting variance in the data, the predicted increases of incongruence effects by such increases of the salience of the meaning of the involved categories are also always proof of the internal validity of the A-CIT.

Finally, a yet unrealized potential is the combination of RT-based A-CIT with other deception-detection methods that use sequentially presented simple stimuli (e.g., polygraph, EEG; Meijer, Smulders, Johnston, & Merckelbach, 2007; Meijer et al., 2016). Using the same or a similar task, the focus on the associations may not only lead to larger differences in RTs, but may also improve the differentiability of the physiological responses to the probe item (see, e.g., Williams & Themanson, 2011).

4.2. Conclusions

The current paper used three Association-based Concealed Information Test versions that successfully distinguished between guilty and innocent participants (simulating guilty and innocent suspects). Out of these three, two were novel versions that both showed promisingly high validity and outperformed the conceptual replication of the original version. We have also described several ways these methods could be further improved. If these further improvements on the A-CIT are as successful as in the current study, this easily and quickly implementable method has the potential to surpass the classification accuracies of such complex high-tech technologies as the EEG or the extremely costly fMRI, and thereby pave the path for real life application.

Acknowledgements

Gáspár Lukács is a recipient of a DOC Fellowship of the Austrian Academy of Sciences at the Institute for Basic Psychological Research at the University of Vienna. The participant fee expenses were reimbursed (after the completion of the study) by a “Förderungsstipendien nach dem StudFG” research grant from the University of Vienna. We thank Bennett Kleinberg for the basis of the LexTALE task JavaScript code, and Dorota Goc and several others for repeatedly pretesting the tasks.

Appendix A. Inducer items and labels

Each version had a set of self-referring inducers. In each block, half of the four inducers were generic (e.g., “mine”), and the other half was specific to the category of probe and irrelevant items presented in that block (countries, dates, or animals; e.g., “my country” for countries). In

the Side-label and Center-label versions, the generic self-referring inducers were “MINE” and “SELF RELATED”. The category-specific self-referring inducers were: “MY HOMELAND” and “OWN COUNTRY” for countries of origin; “MY BIRTHDAY” and “OWN DATE” for birthdays; “MY ANIMAL” and “OWN FAVORITE” for favorite animals. In both these versions, the category labels were “mine” and “other.” In the Side-label version, these were displayed in the bottom left and right corners of the screen, and corresponded to the response keys as well. In the Center-label version, these category labels appeared at the center of the screen. As reminders, the response labels “no” and “yes” were displayed in the bottom left and right corners of the screen (but, as described under the Procedure section, these did not require attention throughout the task).

In the Color-based version, there were also other-referring inducers, again including both generic and category-specific ones. However, to keep the number of different inducers equal, we used only half of the above described self-referring inducers. The same number of other-referring inducers were then added to these, so that altogether the Color-based version had the same number of inducers (with each presented the same number of times) as in the other two versions. Therefore, the only generic self-referring inducer was “MINE”, while the category-specific self-referring inducers were “MY HOMELAND” for countries of origin; “MY BIRTHDAY” for birthdays; and “MY ANIMAL” for favorite animals. The generic other-referring inducer was “OTHER”, while the category-specific other-referring inducers were “OTHER COUNTRY” for countries of origin; “OTHER DATE” for birthdays; and “OTHER ANIMAL” for favorite animals. Here, there were no category labels at all. As reminders, the response labels “not mine” and “mine” were displayed in the bottom left and right corners of the screen (which, again, did not require specific attention; see also under Procedure and footnote 4).

References

Bluemke, M., & Friese, M. (2008). Reliability and validity of the single-target IAT (ST-IAT): Assessing automatic affect towards multiple attitude objects. *European Journal of Social Psychology, 38*(6), 977–997. https://doi.org/10.1002/ejsp.487.

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

Bond, C. F., & DePaulo, B. M. (2008). Individual differences in judging deception: Accuracy and bias. *Psychological Bulletin, 134*(4), 477–492. https://doi.org/10.1037/0033-2909.134.4.477.

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

De Houwer, J., & De Bruycker, E. (2007). The identification-EAST as a valid measure of implicit attitudes toward alcohol-related stimuli. *Journal of Behavior Therapy and Experimental Psychiatry, 38*(2), 133–143. https://doi.org/10.1016/j.jbtep.2006.10.004.

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. https://doi.org/10.1037/a0023589.

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

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

Kraut, R. (1980). Humans as lie detectors. *Journal of Communication, 30*(4), 209–218. 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), https://doi.org/10.3389/fpsyg.2013.00863.

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods, 44*(2), 325–343. https://doi.org/10.3758/s13428-011-0146-0.

Lukács, G., & Ansorge, U. (2018). *Information leakage in the response time-based concealed information test.* (Manuscript submitted).

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

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

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

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

Meissner, F., & Rothermund, K. (2013). Estimating the contributions of associations and recoding in the implicit association test: The ReAL model for the IAT. *Journal of Personality and Social Psychology, 104*(1), 45–69. https://doi.org/10.1037/a0030734.

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. https://doi.org/10.1002/acp.2910.

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science, 7*(6), 657–660. https://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*. https://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), Retrieved from https://archives.fbi.gov/archives/about-us/lab/forensic-science-communications/fsc/july2003/podlesny.htm.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin, 114*(3), 510–532. https://doi.org/10.1037/0033-2909.114.3.510.

Rothermund, K., Teige-Mocigemba, S., Gast, A., & Wentura, D. (2009). Minimizing the influence of recoding in the implicit association test: The recoding-free implicit association test (IAT-RF). *Quarterly Journal of Experimental Psychology, 62*(1), 84–98. https://doi.org/10.1080/17470210701822975.

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. https://doi.org/10.1037/0021-9010.85.1.30.

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

Sriram, N., & Greenwald, A. G. (2009). The brief implicit association test. *Experimental Psychology, 56*(4), 283–294. https://doi.org/10.1027/1618-3169.56.4.283.

Suchotzki, K., Verschuere, B., Van Bockstaele, B., Ben-Shakhar, G., & Crombez, G. (2017). Lying takes time: A meta-analysis on reaction time measures of deception. *Psychological Bulletin, 143*(4), 428–453. https://doi.org/10.1037/bul0000087.

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

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*(7), 991–1002. https://doi.org/10.1002/acp.1601.

Verschuere, B., & Kleinberg, B. (2017). Assessing autobiographical memory: The web-based autobiographical implicit association test. *Memory, 25*(4), 520–530.

Verschuere, B., Kleinberg, B., & Theocharidou, K. (2015). RT-based memory detection: Item saliency effects in the single-probe and the multiple-probe protocol. *Journal of Applied Research in Memory and Cognition, 4*(1), 59–65. https://doi.org/10.1016/j.jarmac.2015.01.001.

Williams, J. K., & Themanson, J. R. (2011). Neural correlates of the implicit association test: Evidence for semantic and emotional processing. *Social Cognitive and Affective Neuroscience, 6*(4), 468–476. https://doi.org/10.1093/scan/nsq065.

Zou, K. H., O'Malley, A. J., & Mauri, L. (2007). Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation, 115*(5), 654–657. https://doi.org/10.1161/CIRCULATIONAHA.105.594929.