Assessing autobiographical memory

The web-based autobiographical Implicit Association Test

Verschuere, B.; Kleinberg, B.

Published in:
Memory

DOI:
10.1080/09658211.2016.1189941

Link to publication

Creative Commons License (see https://creativecommons.org/use-remix/cc-licenses):
CC BY-NC-ND

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Assessing autobiographical memory: the web-based autobiographical Implicit Association Test

Bruno Verschuere & Bennett Kleinberg

To cite this article: Bruno Verschuere & Bennett Kleinberg (2017) Assessing autobiographical memory: the web-based autobiographical Implicit Association Test, Memory, 25:4, 520-530, DOI: 10.1080/09658211.2016.1189941

To link to this article: http://dx.doi.org/10.1080/09658211.2016.1189941

© 2016 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

Published online: 09 Jun 2016.
Assessing autobiographical memory: the web-based autobiographical Implicit Association Test

Bruno Verschuere and Bennett Kleinberg

Department of Clinical Psychology, University of Amsterdam, Amsterdam, The Netherlands

ABSTRACT
By assessing the association strength with TRUE and FALSE, the autobiographical Implicit Association Test (aIAT) [Sartori, G., Agosta, S., Zogmaister, C., Ferrara, S. D., & Castiello, U. (2008). How to accurately detect autobiographical events. Psychological Science, 19, 772–780. doi:10.1111/j.1467-9280.2008.02156.x] aims to determine which of two contrasting statements is true. To efficiently run well-powered aIAT experiments, we propose a web-based aIAT (web-aIAT). Experiment 1 (n = 522) is a web-based replication study of the first published aIAT study [Sartori, G., Agosta, S., Zogmaister, C., Ferrara, S. D., & Castiello, U. (2008). How to accurately detect autobiographical events. Psychological Science, 19, 772–780. doi:10.1111/j.1467-9280.2008.02156.x; Experiment 1]. We conclude that the replication was successful as the web-based aIAT could accurately detect which of two playing cards participants chose (AUC = .88; Hit rate = 81%). In Experiment 2 (n = 424), we investigated whether the use of affirmative versus negative sentences may partly explain the variability in aIAT accuracy findings. The aIAT could detect the chosen card when using affirmative (AUC = .90; Hit rate = 81%), but not when using negative sentences (AUC = .60; Hit rate = 53%). The web-based aIAT seems to be a valuable tool to facilitate aIAT research and may help to further identify moderators of the test's accuracy.

The autobiographical Implicit Association Test (aIAT; Sartori, Agosta, Zogmaister, Ferrara, & Castiello, 2008) is a test for assessing autobiographical memory that has several attractive features. Administration requires a single computer and can be completed in about 10–15 minutes. The potential applications are many, as the aIAT can not only be used to assess autobiographical statements regarding past actions (Sartori et al., 2008), but also of planned actions (Zangrossi, Agosta, Cervesato, Tesarotto, & Sartori, 2015), and even the motives of past and future actions (Agosta, Pezzioli, & Sartori, 2013). Research on its validity, however, is limited and has provided mixed findings (see e.g., Vargo & Petroczi, 2013; Verschuere, Prati, & De Houwer, 2009). Here, we propose a web-based version of the aIAT that will allow efficiently running high-powered aIAT studies and may help to clarify the validity of the aIAT.

The aIAT
The aIAT assesses the veracity of two autobiographical statements. These statements need to be mutually exclusive. A professional cyclist, for instance, could be tested on doping use by contrasting I HAVE USED EPO versus I HAVE NEVER USED EPO. The aIAT determines the veracity of the two statements by comparing the ease of pairing them with TRUE or FALSE.¹ To do so, the aIAT compares performance of two tasks. In both tasks, the examinee classifies one set of sentences (e.g., “I am taking a test”) as belonging to the label TRUE or FALSE, and another set of sentences (e.g., “EPO helped me perform better”) as belonging to the label I HAVE USED EPO or I HAVE NEVER USED EPO. The difference between the tasks is the positioning of the labels. In one task I HAVE USED EPO is combined with TRUE and I HAVE NEVER USED EPO is combined with FALSE. In the other task I HAVE USED EPO is combined with FALSE and I HAVE NEVER USED EPO is combined with TRUE. Participants classify the sentences as belonging to the two categories. Faster and more correct responding to the task where I HAVE USED EPO and TRUE (I HAVE NEVER USED EPO and FALSE) are combined than to the task where I HAVE USED EPO and FALSE (I HAVE NEVER USED EPO and TRUE) are combined would be taken as an indication that the cyclist used EPO. The opposite pattern is expected in a cyclist who never used doping (see Table 1 for a detailed structure of the aIAT).

¹ © 2016 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group. This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.
Initial validation research

In the initial validation study, Sartori et al. (2008) reported upon a series of experiments. In Experiment 1, participants selected one of two playing cards: Seven of clubs or four of diamonds. Participants (n = 20) who had chosen the seven of clubs were faster in the task pairing I HAVE THE 7 with TRUE (I HAVE THE 4 with FALSE) than in the task pairing I HAVE THE 7 with FALSE (I HAVE THE 4 with TRUE), whereas participants (n = 17) who had chosen the four of diamonds were faster in the task pairing I HAVE THE 4 with TRUE (I HAVE THE 7 with FALSE) than in the task pairing I HAVE THE 4 with FALSE (I HAVE THE 7 with TRUE). The differential performance in the two tasks was summarised for each participant in a standardised difference score, the D score (roughly comparable to Cohen’s d; Cohen, 1988; and now routinely used in IAT research see Greenwald et al., 2003). Based upon the D score, the aIAT could correctly indicate for 35 out of 37 (95%) participants which card they had chosen. In Experiment 2, Sartori and colleagues used a mock-crime paradigm with “guilty” participants (n = 15) stealing a CD and “innocent” participants (n = 15) reading a newspaper article on the mock CD theft. The results also supported the aIAT assumptions, and all guilty (100%) and 13 out of 15 (87%) innocents were correctly classified resulting in an overall accuracy of 93%. Four more experiments in the initial validation study as well as further studies by the developers of the aIAT (Agosta, Castiello, Rigoni, Lionetti, & Sartori, 2011; Agosta, Ghirardi, Zogmaister, Castiello, & Sartori, 2011; Agosta, Mega, & Sartori, 2011; Agosta et al., 2013; Zangrossi et al., 2015) yielded very high accuracy.

Independent validation research

While a number of independent studies have also found that the aIAT can reach high accuracy in determining which of two autobiographical statements is true (e.g., Hu & Rosenfeld, 2012; Hu, Rosenfeld, & Bodenhausen, 2012; Takarangi, Strabge, Shortland, & James, 2013), other studies found lower accuracy rates (Morgan, Rabinowitz, Leidy, & Coric, 2014; Vargo & Petroczi, 2013; Verschuere et al., 2009; for reviews of aIAT studies see Agosta & Sartori, 2013; Verschuere, Suchotzki, & Debey, 2015). Administering an aIAT to detect cocaine use, Vargo and Petroczi (2013) found that the aIAT correctly detected 78% of 23 recreational cocaine users, but misclassified 61% of 23 non-cocaine users. It should be noted that the methodology of the aIAT differed substantially from that proposed by the developers (i.e., much shorter task, and stimulus presentation and RT recording through a touch screen). Verschuere et al., 2009 (control conditions of Experiment 1, n = 36, and Experiment 3, n = 42) conducted a close replication of Sartori et al.’s (2008) Experiment 2. While support for the aIAT’s assumptions was found, the obtained classification accuracy (67–87% of guilty and 61% innocent participants) appears lower than that obtained by Sartori et al. (2008). Two questions arise regarding the validity of the aIAT: To what extent is the

- Table 1. Blocks with stimuli for Experiment 1 (A/B protocol).

<table>
<thead>
<tr>
<th>Block number</th>
<th>Left label</th>
<th>Right label</th>
<th>Example stimulus</th>
<th>Number of trials</th>
<th>Passing criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TRUE</td>
<td>FALSE</td>
<td>“I am climbing a mountain” (false); “I am inside” (true)</td>
<td>20</td>
<td>- Mean RT ≤ 5000 ms - Error rate ≤ .3 - Too fast trials ≤ .2 - Too slow trials ≤ .2</td>
</tr>
<tr>
<td>2 (5)</td>
<td>I TURNED OVER THE FOUR OF DIAMONDS</td>
<td>I TURNED OVER THE SEVEN OF SPADES</td>
<td>“I took the seven of spades”; “I saw the four of diamonds”</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>3 (6)</td>
<td>TRUE OR I TURNED OVER THE FOUR OF DIAMONDS</td>
<td>FALSE OR I TURNED OVER THE SEVEN OF SPADES</td>
<td>“I am sitting on the beach”; “I am looking at a computer screen”; “I turned over the seven of spades”; “I got the four of diamonds”</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>4 (7)</td>
<td>Same as block 3</td>
<td>I TURNED OVER THE SEVEN OF SPADES</td>
<td>“I took the seven of spades”; “I saw the four of diamonds”</td>
<td>40</td>
<td>No feedback – test block - Mean RT ≤ 5000 ms - Error rate ≤ .3 - Too fast trials ≤ .2 - Too slow trials ≤ .2</td>
</tr>
<tr>
<td>5 (2)</td>
<td>I TURNED OVER THE SEVEN OF SPADES</td>
<td>I TURNED OVER THE FOUR OF DIAMONDS</td>
<td>“I am sitting on the beach”; “I am looking at a computer screen”; “I turned over the seven of spades”; “I got the four of diamonds”</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>6 (3)</td>
<td>TRUE OR I TURNED OVER THE SEVEN OF SPADES</td>
<td>FALSE OR I TURNED OVER THE FOUR OF DIAMONDS</td>
<td>“I am sitting on the beach”; “I am looking at a computer screen”; “I turned over the seven of spades”; “I got the four of diamonds”</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>7 (4)</td>
<td>Same as block 6</td>
<td></td>
<td></td>
<td>40</td>
<td>No feedback – test block</td>
</tr>
</tbody>
</table>
alAT effect replicable? and Are there moderators that can explain variability in the alAT effect? These questions fit the current concerns regarding the replicability of psychological research, and a growing awareness of the importance of replication research (Open Science Collaboration, 2015).

The current study

We introduce a new, web-based research tool for the alAT. A notable benefit of web-based research is the efficiency to run high-powered studies. Using this web-based alAT (web-alAT), we first conducted a replication study of the card game (Sartori et al., 2008; Experiment 1). Next, we tested the use of affirmative versus negative sentences as a potential moderator to explain variability in alAT accuracy findings (Agosta, Mega et al., 2011). Agosta and colleagues showed that alAT studies using affirmative sentences (e.g., I AM SOBER versus I AM DRUNK) showed high validity and alAT studies using negative sentences (e.g., I AM SOBER versus I AM NOT SOBER) showed lower or more variable validity, but these authors did not experimentally manipulate affirmative versus negative sentences. Interestingly, most studies that have found moderate accuracy (Morgan et al., 2014; Verschuere et al., 2009) did indeed use negative sentences, indirectly supporting the idea that negative sentences are detrimental to the alAT. In the only direct comparison of affirmative versus negative sentences, Vargo and Petroczi (2013) failed to find differences between the two protocols. As mentioned above, however, their methodology differed substantially from that of the typical alAT. Using the web-based alAT, we thus not only examine the replicability of the alAT (Experiment 1; and affirmative sentences condition of Experiment 2), but provide a well-powered direct empirical test of a possible moderator (negative versus affirmative sentences; Experiment 2).

Pilots

We ran several small pilot studies to debug our programme.

Experiment 1

Experiment 1 is a web-based replication study of the first published alAT study (Sartori et al., 2008; Experiment 1). Participants chose one of two playing cards (seven of spades or four of diamonds). By examining the ease of associating sentences related to choosing the seven of spades or four of diamonds with TRUE and FALSE, the alAT aimed to determine the chosen card.

Method

The experimental task is accessible via this link: http://www.liereresearch.com/?page_id=704. Experiments 1 and 2 have been approved by the IRB of the Department of Clinical Psychology of the University of Amsterdam (2014-CP-3945). All data are publicly available via the OSF repository https://osf.io/swp27/.

Participants

The experiment was advertised as a task on the online workforce platform CrowdFlower (a crowdsourcing platform similar to Amazon Mechanical Turk). Each participant was offered $0.50 as compensation. We used no in- or exclusion criteria on CrowdFlower based on previous performance on CrowdFlower tasks. We aimed for n = 250, but due to great number of participants starting simultaneously, ended up with n = 522. We excluded participants based on the following preliminary criteria: not completing all seven blocks (preventing calculation of key results; n = 36); IP address being recorded more than once (indicating possible multiple participation; n = 15); use of translation function of browser (resulting in inaccurate timing and RT measurement; n = 28); age lower than 18 (as we planned to test adults only) see also Verschuere, Kleinberg, & Theocharidou, 2015; n = 5); incorrect post-test recall of the selected card (indicating lack of memorisation of the card; n = 1). The final sample consisted of 437 participants (overall data loss: 16%), of which 28% were female. The overall mean age was 31 years (SD = 9). More than 40 different native languages were reported, with the most common native language being English (14%). Sixty-two different countries of origin were reported, with the most common country being India (13.5%). Participants had been randomly assigned to the “seven of spades” (n = 210, 29.5% female, M_age = 31, SD_age = 9) or the “four of diamonds” condition (n = 227, 26% female, M_age = 31, SD_age = 9). The conditions did not differ in gender, χ^2(1) = 0.38, p = .540, or age, F(1, 435) = 0.01, p = .928, f < 0.01.

Procedure

Upon beginning with the task (offered in Dutch and English), each participant went through this experiment in their own pace on their own computer (i.e., a desktop or laptop, but not on a smartphone or IPad). Each participant gave written informed consent to this task and started by indicating demographic details (age, gender, country of origin, native language). Hereafter, they chose and memorised one of two playing cards (http://www.liereresearch.com/wp-content/uploads/AlAT02/en/page05.html; see) and passed three checks with regard to their card (http://www.liereresearch.com/wp-content/uploads/AlAT02/en/page06.html). After selecting a number of sentences as being true or false (http://www.liereresearch.com/wp-content/uploads/AlAT02/en/stim_sel.html), they were directed to the first of seven aLAT blocks (http://www.liereresearch.com/wp-content/uploads/AlAT02/en/cb1/block1.html). After completing the aLAT, knowledge of the
chosen card was checked again and subjective ratings were completed after which they received their result ("Based on your reaction times, the test guesses that you have turned over the [seven of spades/four of diamonds]"). After submitting their unique participation codes, the participants were re-directed to the CrowdFlower starting page.

Card game

The participant chose one of two playing cards by clicking on the neutral backside of one card, upon which the card turned and showed the participants the identity of the card. Participants were instructed to memorise their chosen card (for as long as they wanted). Below the turned-over card was a sentence stating their choice ("You selected the [seven of spades/four of diamonds]"). The card was pre-defined to be either the four of diamonds or the seven of spades. Next, a forced choice menu appeared with pictures of eight possible playing cards (four black and four red) one of which was the chosen card. The participants were told to select their card and received positive feedback if their choice was correct and negative feedback if not. If they misidentified their card, they had to choose again until they correctly chose their card. Participants went through this card check three times, each of which with different order of the cards. Derived from this card game, five statements for the I TURNED OVER THE FOUR OF DIAMONDS category (i.e., "I clicked on the four of diamonds", "I got the four of diamonds", "I saw the four of diamonds", "I turned over the four of diamonds", and "I took the four of diamonds"), and five statements for the I TURNED OVER THE SEVEN OF SPADES category (i.e., "I clicked on the seven of spades", "I got the seven of spades", "I saw the seven of spades", "I turned over the seven of spades", and "I took the seven of spades") were pre-defined as stimuli for the aIAT.

True and false statements

Sartori et al. (2008) used autobiographical statements that the experimenter could verify as being true (e.g., "I'm in the psychology laboratory") or false (e.g., "I'm eating in a downtown restaurant"). As we could not know where participants were and what they were doing, we asked them to indicate true and false autobiographical statements from a list. First, participants selected at least five out of ten possible statements that were true for them at that time (i.e., "I am sitting on a chair", "I am sitting on a couch", "I am looking at a computer screen", "We live in the year 2014", "I am taking a test", "There is a keyboard in front of me", "I am dressed", "I am inside", "I am online", and "I am participating in a research project"). They could select and deselect the sentences by clicking on them and were only able to proceed with at least five options selected. After they selected the true statements, they did the same for presumably false statements (i.e., "I am playing football", "I am climbing a mountain", "We live in the year 2008", "I am taking a bath", "I am watching television", "I am sitting on a beach", "I am walking in the forest", "I am wearing pajamas", "I am outside", and "I am cooking dinner"). If more than five statements were selected in either category, five of these statements were selected randomly for further use in the experiment.

Autobiographical IAT

The autobiographical IAT consisted of seven blocks (see Table 1). The inter-stimulus interval between two stimuli was completely random 250, 500 or 750 ms. There was no response deadline so the stimuli remained in the centre of the screen until the participant responded with a key press. The e-key and the i-key were the required response keys to categorise the stimuli. If the categorisation was incorrect (e.g., categorising a false sentence like "I am sitting on the beach" to TRUE), a red X appeared below the stimulus until the incorrect response was corrected. For each trial, reaction time (RT) was defined as the elapsed time between the presentation of the stimulus and the correct key press. For all but the test blocks (5 and 7), participants had to meet a number of criteria in order to proceed to the next block (see also Kleinberg & Verschuere, 2015). These passing criteria were implemented to ensure that participants understood the task and did not rush through it. The specific criteria were (1) the overall mean RT must not exceed 5000 ms; (2) the percentage of incorrect responses per block must not exceed 30%; (3) the percentage of trials per block with an RT smaller than 300 ms must not exceed 20%; and (4) the percentage of trials per block with an RT larger than 10,000 ms must not exceed 20%. The latter two criteria were chosen to indicate responses that were too fast and too slow, respectively, for genuinely processed trials (Table 1). If one or more criteria were not met, the participant received feedback in a pop-up pointing out the reason for the feedback (e.g., "You are making too many mistakes") and requiring the participant to repeat the respective block. Within each block, the stimuli were presented twice in a random sequence.

The first block presented the labels TRUE and FALSE in yellow font in the upper left and rights part of the screen, respectively. The stimuli used for the true/false discrimination were those that were selected by the participant (e.g., "There is a keyboard in front of me", "I am sitting on the beach"). These stimuli appeared in the centre of the screen in yellow fixed-width font against a black background. Participants pressed left (e-key) for TRUE and right (i-key) for FALSE sentences.

The second block consisted of the categorisation of sentences about playing cards into the categories I TURNED OVER THE FOUR OF DIAMONDS (press left; e-key) or I TURNED OVER THE SEVEN OF SPADES (press right; i-key).
The card-labels and the card-stimuli were displayed in white colour.

The third block combined the labels of the previous two blocks, whereby the labels were TRUE or I TURNED OVER THE FOUR OF DIAMONDS and FALSE or I TURNED OVER THE SEVEN OF SPADES (http://www.lieresearch.com/wp-content/uploads/AIAT02/en/cb1/block3p.html). In order to facilitate the categorisation, the label colours remained consistent throughout, that is the true/false label and stimuli were yellow, and the card-labels and stimuli were white. The task in this block was to categorise each stimulus into one of the two combined labels (e.g., “I got the four of diamonds” and “I am inside” were both categorised to the left label). The third block served as practice (20 trials) for the fourth block, which was identical to the third block, but with twice as many trials (40 trials).

The fifth block required the participants to categorise the card-stimuli to the same labels as in block 2 but with the categories on the opposite sides (I TURNED OVER THE FOUR OF DIAMONDS left versus I TURNED OVER THE SEVEN OF SPADES right). This blocked prepared participants for the reversed combination in blocks 6 and 7.

The sixth and seventh blocks (http://www.lieresearch.com/wp-content/uploads/AIAT02/en/cb1/block5p.html) were similar to blocks 3 and 5, with the critical difference that the card-labels had switched sides (i.e., TRUE or I TURNED OVER THE SEVEN OF SPADES and FALSE or I TURNED OVER THE FOUR OF DIAMONDS).

The order of blocks was counterbalanced so that the I TURNED OVER THE FOUR OF DIAMONDS was first combined with TRUE for some participants and with FALSE for the other participants (see Table 1, column 1).

Upon completion of the seventh block, participants were presented the same eight playing cards as in the card scenario in random order and had to select the card they chose at the beginning of the experiment.

As final step of the experimental task, participants received their individual result (see⁴). Participants were also provided a unique participation code, which they had to submit alongside their CrowdFlower participation number in order to verify their participation for payment purposes.

Results

Analysis plan

For all analyses, the alpha level of .05 was chosen. Cohen’s f effect sizes (Cohen, 1988) are reported for F-tests, and Cohen’s d effect sizes are reported for follow-up t-tests, whereby we differentiate between dwithin and dbetween (Lakens, 2013) for within-subjects factors and between-subjects comparisons, respectively. For one-sampled t-tests with μ = 0, we used dz = t/sqrt(n) (Lakens, 2013). The main analysis for Experiment 1 was a one-way ANOVA with Card (seven of spades versus four of diamonds) as the between-subjects factor and D transformed IAT score as the dependent variable.⁵ We calculated the DIAT score (Greenwald, Nosek, & Banaji, 2003) as follows: we (i) excluded RTs larger than 10,000 ms, (b) excluded participants with more than 10% of trials faster than 300 ms (no exclusion), (c) used the RTs as per correct response (i.e., the second RT per trial, see above), (d) we recoded the block order so that block 3 and 4 represent the association between TRUE and SEVEN OF SPADES (FALSE and FOUR OF DIAMONDS) for all participants and block 6 and 7 that between TRUE and FOUR OF DIAMONDS (FALSE and SEVEN OF SPADES), (e) calculated mean RTs for block 3, 4, 6, and 7, (f) calculated the pooled SDs for block 3 and 6, and for block 4 and 7, (g) calculated the RT difference for blocks 4–7 and for blocks 3–6, (h) divided the RT difference by the respective pooled SD, and (i) averaged the two resulting composite scores to one D score. The resulting D score was used for further analysis. A positive D score indicates a stronger association between TRUE and FOUR OF DIAMONDS (FALSE and SEVEN OF SPADES) than between FALSE and FOUR OF DIAMONDS (TRUE and SEVEN OF SPADES). A negative score indicates a stronger association between TRUE and SEVEN OF SPADES (FALSE and FOUR OF DIAMONDS) than between FALSE and SEVEN OF SPADES (TRUE and FOUR OF DIAMONDS).

Main analysis

The one-way ANOVA with Card (seven of spades versus four of diamonds, between-subjects) on the D scores revealed a significant main effect of Card, F(1, 435) = 316.17, p < .001, f = 0.85 showing that the D scores differed according to the chosen card. One-sample t-tests against zero indicated that the D score for participants with seven of spades (M = −0.34, SD = 0.47) was significantly negative, t(209) = 10.68, p < .001, dz = 0.74 [95% Confidence interval: 0.58–0.89], and that the D score for participants with four of diamonds (M = 0.44, SD = 0.45) was significantly positive, t(226) = 14.59, p < .001, dz = 0.97 [0.81–1.13].

Figure 1. Mean RTs for the block-pairing TRUE with I TURNED OVER FOUR OF DIAMONDS versus the block pairing TRUE with I TURNED OVER SEVEN OF SPADES for participants who picked the four of diamonds versus the seven of spades in Experiment 1 (with standard errors).
To illustrate, Figure 1 depicts mean RTs. Aggregated across chosen card, the RTs (calculated similarly as in the D measure) for the block in which the chosen card was combined with true (and non-chosen card with false; $M = 1317, SD = 382$) were significantly faster than for the block in which the chosen card was combined with false (and non-chosen card with true; $M = 1647, SD = 548$), $t(436) = 14.89, p < .001, d_{within} = 0.71 [0.61–0.82]$. In order to test how well the $D$ measure discriminated between participants having chosen seven of diamonds versus four of diamonds, we examined Receiver Operating Characteristics (ROC) and calculated the area under the curve (AUC; Hanley & McNeill, 1983). Values close to 1.0 indicate better classification performance whereas a value of 0.5 equals random classification. The AUC was significantly above chance at $0.88 [0.85–0.91]$.

Following Sartori et al. (2008), we classified participants with a negative $D$ score ($D > 0$) as having chosen seven of spades and participants with a positive $D$ scores ($D < 0$) as having chosen four of diamonds. This classification rule correctly classified 353 of 437 (81%) participants. This classification rule, however, is not necessarily the optimal cut-off, and we explored two alternatives. First, Noordraven and Verschuere (2013) as well as Agosta and Sartori (2013) noted that it may be arbitrary to classify participants with $D$ scores around zero (e.g., classifying a participant with a $D$ score of $+0.001$), and suggested to have an inconclusive category for participants with $D$ scores around zero ($-0.2 < D < +0.2$). Thus, only participants with an absolute $D$ score of at least 0.2 (positive or negative) were classified. This excluded 99 participants, and resulted in correct classification 287 of 338 (85%) participants. Using the inconclusive category slightly improved accuracy, but at the cost of having a substantial amount of unclassified participants (23%). Second, based upon the Youden’s $J$ statistic (Robin et al., 2011; Youden, 1950), we derived the “optimal” cut-off point by calculating the $D$ score that allowed for the best classification. Because such an optimal cut-off point (here: $D = 0.13$) capitalises on chance, we split our sample (odd–even split) semi-randomly in half, with one half (“model-building sample”; $n = 218$) used to calculate Youden’s $J$ statistic, and the other half (“validation sample”; $n = 219$) for validation. In the model-building sample 174 of 218 (80%) participants were correctly classified, and in the validation sample 173 of 219 (79%) participants were correctly classified. Classification based upon Youden’s $J$ thus did not improve accuracy.

**Discussion**

Experiment 1 presents the first study to validate the web-based aIAT using a close replication of Sartori et al.’s (2008) Experiment 1. There are some notable differences in methodology between our study design and that of Sartori et al.’s (2008) Experiment 1. We used (1) an online test, with (2) extensive exclusion and passing criteria to guarantee high-quality data, in (3) an ethnically diverse sample, (4) testing mostly in non-native language, (5) using participant-selected true and false sentences, (6) with a short pause between practice and test blocks, and (7) 20 trials in block 5 (instead of 40) to prepare for the blocks with the reversed pairings. Despite those differences, our result are similar to those of Sartori et al. Participants were faster to categorise sentences corresponding to their chosen card and true and sentences corresponding to the non-chosen card and false. In Experiment 2, we used the same card game to test a potential moderator of the aIAT’s accuracy: Affirmative versus negative sentences. Participants of the card game were tested with an aIAT consisting of affirmative ( = replication of Experiment 1) or negative sentences.

**Experiment 2**

The experimental task is accessible via this link: [http://www.lieresearch.com/?page_id=716](http://www.lieresearch.com/?page_id=716)

There was only one key change to the design of Experiment 1: The addition of a between-subjects factor test protocol. In Experiment 1 the labels (I TURNED OVER THE FOUR OF DIAMONDS versus I DID NOT TURN OVER THE FOUR OF DIAMONDS) were phrased affirmatively. This test protocol has also been labelled the A/B protocol (Verschuere, Suchotzki, et al., 2015). In Experiment 2 we added the A/not-A version, in which one category is a simple negation of the other category (I TURNED OVER THE FOUR OF DIAMONDS versus I DID NOT TURN OVER THE FOUR OF DIAMONDS). We phrased all sentences of the A/not-A version negatively (i.e., “I did not click on the four of diamonds”, “Four of diamonds was not the card I got”, “I didn’t see the four of diamonds”, “I didn’t turn over the four of diamonds”, and “Four of diamonds was not the card I took”). While the negative phrasing could be manipulated independently for labels and statements (Agosta et al., 2011), we chose to maximise the effects by phrasing both the label and the sentences of the contrast category negatively. The resulting design for Experiment 2 was thus 2 (Card: four of diamonds versus seven of spades) by 2 (Protocol: A/B versus A/not-A).

A power analysis using G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) for a $2 \times 2$ ANOVA with $\alpha = .05$, and $\beta = .80$, showed that 128 participants are needed to be able to pick up the predicted interaction effect of at least moderate size $f = .25$.
Participants
The data were collected again via CrowdFlower but this time with an exclusion criterion that only trusted participants were allowed to participate. Specifically, we set this criterion so that only those participants who successfully completed at least 20% of previous tasks on CrowdFlower were allowed. Participants received $0.50 as compensation. We aimed for \( n = 400 \), as to have at least 100 participants in each of the 4 conditions of our 2 (four of diamonds versus seven of spades) × 2 (A/B protocol versus A/not-A protocol) design. Due to simultaneous starting times, the sample consisted of 424 participants. Forty-one participants were excluded for not completing all blocks; 23 were excluded based on double IP addresses; 12 participants were excluded because they translated parts of the task; 2 were excluded for indicating an age lower than 18; and 3 participants were excluded for not recalling the selected card. The final sample consisted of 343 participants (overall data loss: 19%). Participants reported more than 33 different native languages, with the most commonly reported native language being English (17%). Participants reported 64 different countries of origin with the most common countries of origin being India (13%). Participants had been assigned to one of the four conditions: seven of spades and A/B \( (n = 88, 26\% \text{ female}, M_{\text{age}} = 31, SD_{\text{age}} = 9) \), seven of spades and A/not-A \( (n = 88, 28\% \text{ female}, M_{\text{age}} = 32, SD_{\text{age}} = 11) \), four of diamonds and A/B \( (n = 85, 29\% \text{ female}, M_{\text{age}} = 31, SD_{\text{age}} = 10) \), four of diamonds and A/not-A \( (n = 82, 27\% \text{ female}, M_{\text{age}} = 30, SD_{\text{age}} = 8) \). The four groups did not differ in gender, \( X^2(3) = 0.29, p = .967 \), or age, \( F(1, 341) = 0.58, p = .447, f = 0.04 \).

Results

Main analysis
The 2 (Card: seven of spades vs four of diamonds, between-subjects) by 2 (Protocol: A/B vs A/not-A, between-subjects) ANOVA on the D scores revealed a significant main effect of Card, \( F(1, 339) = 122.71, p < .001, f = 0.60 \); a significant main effect of Protocol, \( F(1, 339) = 140.25, p < .001, f = 0.64 \), and a significant interaction between Card and Protocol, \( F(1, 339) = 62.21, p < .001, f = 0.43 \). This interaction indicated that the D scores displayed the expected pattern for the A/B protocol but not for the A/not-A protocol.

For the A/B protocol, as in Experiment 1, the D scores differed significantly with the chosen card, \( t(166.17) = 12.90, p < .001 \), \( d_{\text{between}} = 1.96 [1.59–2.32] \). Participants who chose seven of spades \( (M = –0.40, SD = 0.49) \) showed a significant negative D score, \( t(87) = 7.63, p < .001, d_x = 0.81 [0.57–1.05] \), and the D score for participants who chose four of diamonds \( (M = 0.47, SD = 0.40) \) was significantly positive, \( t(85) = 11.01, p < .001, d_x = 1.19 [0.91–1.47] \).

For the A/not-A protocol, however, the mean D scores were not only significantly positive for participants with the four of diamonds \( (M = 0.66, SD = 0.36), t(80) = 16.31, p < .001, d_x = 1.81 [1.45–2.16] \), but also, and contrasting the aIAT assumptions, for participants who chose the seven of spades \( (M = 0.51, SD = 0.44), t(88) = 11.04, p < .001, d_x = 1.17 [0.90–1.44] \). For sake of completion, we mention that the D score for those with four of diamonds was slightly more positive than for those with seven of spades, \( t(165.60) = 2.38, p = .018, d_{\text{between}} = 0.37 [0.06–0.67] \).

Figure 2 illustrates that mean RTs confirm the aIAT test assumptions for the A/B protocol, but not the A/not-A protocol. For the A/B protocol, the RTs (calculated similarly as in the D measure) for the block in which the chosen card was combined with TRUE (and non-chosen card with FALSE; \( M = 1291, SD = 376 \)) were significantly faster than for the block in which the chosen card was combined with FALSE (and non-chosen card with TRUE; \( M = 1634, SD = 527 \), \( t(172) = 10.54, p < .001, d_{\text{within}} = 0.80 [0.63–0.97] \). For the A/not-A protocol, the mean RT for the block in which the chosen card was combined with TRUE (and non-chosen card with FALSE; \( M = 1653, SD = 499 \)) did not differ from the RTs for the block in which the chosen card was combined with FALSE (and non-chosen card with \( TRUE \) \( (M = 1715, SD = 539), t(169) = 1.37, p = .17, d_{\text{within}} = 0.11 [–0.05–0.26] \).
The ROC analysis also confirmed that the diagnosticity of the A/B protocol (AUC = 0.90; 0.86–0.95) was higher than that of the A/not-A protocol (AUC = 0.60; 0.51–0.68). DeLong’s test for two ROC curves (Robin et al., 2011) revealed that the A/B protocol outperformed the A/not-A protocol significantly, $D$(260.46) = 6.19, $p < .001$.

Likewise, the Hit rates show the same superior performance of the A/B protocol as compared to the A/not-A protocol. With $D = 0$ as cut-off, the Hit rate was 140 out of 173 (81%) for the A/B protocol versus 90 out of 170 (53%) for the A/not-A protocol. With the optimal cut-off determined in Experiment 1 ($D = -0.13$), the Hit rate was 140 out of 173 (81%) for the A/B protocol versus 94 out of 170 (55%) for the A/not-A protocol. When excluding the inconclusive category ($n = 28$ inconclusive test results for the A/B protocol; $n = 21$ inconclusive test results for the A/not-A protocol), the Hit rate was 129 out of 145 (89%) for the A/B protocol versus 81 out of 149 (54%) for the A/not-A protocol. In sum, the A/B protocol systematically outperformed the A/not-A protocol. Having an inconclusive category, but not classification based upon the Youden $J$, seemed to improve accuracy for the A/B protocol, but at the cost of having a substantial amount of unclassified participants (16%).

**Split-half reliability**

As in Experiment 1, we calculated the split-half reliability with Spearman-Brown correction. For the A/B protocol, the reliability was very high, $\rho = .91$ [.88–.93]. For the A/not-A protocol, the reliability was also high, $\rho = .75$ [.69–.82]. Agreement in classification based upon the two halves was 87.28% for the A/B protocol, and 91.76% for the A/not-A protocol. In other words, the A/B protocol is reliably valid, and the A/not-A protocol is reliably invalid.

**General discussion**

The aIAT is highly flexible test of autobiographical memory that requires minimal time, equipment, and preparation for administration. While some studies reported very high accuracy (> 90%; Agosta Castiello, et al., 2011; Agosta, Ghirardi, et al., 2011; Agosta, Mega, et al., 2011; Agosta, et al., 2013; Sartori et al., 2008), others have found lower accuracy (e.g., Verschueren et al., 2009) or even performance at chance level (e.g., Morgan et al., 2014; Vargo & Petroczi, 2013). This led to two research questions: To what extent is the aIAT effect replicable? and Are there moderators of the aIAT effect? To facilitate research on these important questions, we developed a web-based aIAT. Experiment 1 was an online replication study of the first laboratory-based aIAT study. Experiment 2 also had this replication component, but additionally investigated a potential moderator of aIAT accuracy: the use of affirmative versus negative sentences.

**To what extent is the aIAT effect replicable?**

Experiment 1 and the A/B condition of Experiment 2 are close replication studies of the card study by Sartori et al. (2008). The Open Science Collaboration (2015) proposed guidelines to evaluate the reproducibility of a study through a replication attempt: (1) Using significance and $p$-values of the original and the replication study, (2) using effect sizes to compare original and replication study, (3) using subject assessments of the replicating researchers, and (4) conducting a meta-analysis of the effect sizes. Given the limited number of studies, we will rely on the first three guidelines to evaluate replication success. With regard to significance and $p$-values, both Experiments 1 and 2 (A/B condition) were successful replications. The aIAT effect was in the right direction – participants found it easier to pair their chosen card with TRUE (and the non-chosen card with FALSE) than with FALSE (and their non-chosen card with TRUE) – and significant with $p$-values <.001. The effect size in the original study was Cohen’s $d = 1.00$ (95% CI 0.61–1.40; Sartori, personal communication, 17 March 2014). The effect size in the current experiments was Cohen’s $d = 0.71$ [0.61–0.82] for Experiment 1 and Cohen’s $d = 0.80$ [0.63–0.97] for Experiment 2 (A/B condition). These effect sizes fall within the 95% CI of the effect size of the original study. Regarding the subjective assessment, we reasoned that while neither of the current experiments showed the extreme high accuracy reported by Sartori (AUC = .99; Hit rate = 95%), both current experiments showed high accuracy. We therefore answer affirmatively to “Did your results replicate the original effect?” Taken together, we conclude that both Experiment 1 and Experiment 2 (A/B condition) replicate the key finding of Sartori et al. (2008; Experiment 1), and that the card game aIAT effect observed in the lab seems replicable in an online format. An important addition is that this conclusion rests on setting strict exclusion (data loss: 16–19% of participants) and passing criteria for the web-aIAT (see Table 1). The replication success implies that the aIAT findings for the card game generalise to an ethnically diverse sample, tested in non-native language, with participant-selected TRUE and FALSE sentences and an aIAT test format that differs slightly from that used by Sartori et al (see Discussion of Experiment 1).

**Are there moderators that can explain variability in aIAT accuracy?**

In the present study we focused on one possible moderator of the aIAT effect: the use of affirmative (A/B protocol) versus negative sentences (A/non-A) protocol. The developers of the aIAT initially used both protocols interchangeably, but later observed between studies that the A/B protocol more consistently produced high accuracy as compared to the A/non-A protocol (Agosta et al., 2011). One possible reason may be that the A/non-A protocol is more influenced by factors other than veracity. To
determine which of two statements is true, the aIAT relies on the ease of associating the statements with TRUE and FALSE. True autobiographical statements are more easily associated with TRUE and false autobiographical statements are more easily associated with FALSE. However, dimensions other than veracity may also affect the ease of the two tasks (Rothermund & Wentura, 2004). TRUE has a positive meaning and FALSE has a negative meaning. Irrespective of veracity, pairing TRUE with a positive statement (FALSE with a negative statement) will be easier than pairing TRUE with a negative statement (FALSE with a positive statement). Therefore, one should refrain from using labels such as GUILTY versus INNOCENT (see e.g., Vargo, Petroczi, Shah, & Naughton, 2014) because it can be expected that all examinees – irrespective of guilt – will find it easier to do the task in which the labels of the same valence are paired (GUILTY with FALSE; INNOCENT with TRUE) compared to the task where the labels are of different valence (GUILTY with TRUE; INNOCENT with FALSE). The A/not-A version may suffer from the same problem: All participants will find it easier to do the task in which I TURNED OVER FOUR OF DIAMONDS is paired with TRUE (I DID NOT TURN OVER FOUR OF DIAMONDS with FALSE) than the task in which I TURNED OVER FOUR OF DIAMONDS is paired with FALSE (I DID NOT TURN OVER FOUR OF DIAMONDS with TRUE) because in the former task the labels are more similar (both affirmative or both negative) than in the latter task (one affirmative, one negative). While we anticipated such an effect to have some influence on aIAT accuracy, the effects were in fact dramatic. The high accuracy observed in the A/B protocol in Experiments 1–2 (AUC = .88–.90) dropped to being just above chance level (AUC = .60). These findings support the reasoning of Agosta et al. (2011) that negative labels and sentences may be detrimental for aIAT accuracy. Our findings contrast with those of Vargo and Petroczi (2013) who found no effect of affirmative versus negative sentences on the aIAT. As explained above, their design differed substantially from that of the typical aIAT. Moreover, our sample size (n = 424) was substantially larger than theirs (n = 46). Taken together we conclude that the use of negative labels and sentences is an important moderator of the aIAT effect and should be avoided.

Limitations

This study is not without its limitations.

First, we do not have a direct comparison of our web-based aIAT with a laboratory aIAT. Researchers have expressed concern that, due to increased noise, online research may underestimate RT-effects, and that such effects are likely to be larger in laboratory settings. A growing set of empirical findings, however, indicates that such concerns are largely unwarranted (see e.g., Crump, McDonnell, & Gureckis, 2013; Germine et al., 2012). Hilbig (2016), for instance, assigned participants randomly to perform a lexical decision task either online, in the lab using traditional laboratory software, or in the lab using the online test. RT effects were large in all three conditions, with no indication that the online test would result in a lower RT effect.

Second, as in most aIAT studies (for a detailed review see Suchotzki, Verschuere, Van Bockstaele, Ben-Shakhar, & Crombez, 2016), participants were not instructed to hide the true autobiographical event (here: the chosen card), and indeed indicated no motivation to avoid detection or faking attempts. In many applied settings, however, deceptive examinees are likely to be motivated to avoid detection and engage in strategic attempts to alter the test outcome in their favour. To serve as a lie detection tool, it is important to map not only how informed strategic faking affects the outcome of the aIAT (Verschuere et al., 2009), but also whether intention to hide and associated incentives to appear innocent are sufficient to moderate the aIAT effect.

Third, the current studies relied upon a single paradigm, the card game. It will be important to examine whether the findings for the web-based aIAT also hold for other applications of the aIAT (e.g., mock-crime paradigm).

Conclusions

Experiment 1 replicated the key effect of the laboratory-based card aIAT, and showed that high validity can be obtained with the web-aIAT. Experiment 2 showed that subtle differences in methodology (i.e., affirmative versus negative sentences) can dramatically impact upon the aIAT’s validity. The web-aIAT seems a promising new tool to efficiently run high-powered aIAT studies, thereby clarifying the validity of the aIAT and its boundary conditions (Hu et al., 2012; Rothermund & Wentura, 2004; Takarangi et al., 2013; Verschuere et al., 2009).

Notes

1. Throughout the article we put the aIAT category labels in CAPITAL and the items in “quotes”.
2. These links are provided for illustration purposes. To re-run the experiment as the participants did, one needs to start the task from the beginning (e.g., for Experiment 1 start http://www.lieresearch.com/?page_id=704).
3. The passing criteria led to the following passing proportions. A/B: Block 1 was completed in one attempt by 96.10%; block 2: 93.50%; block 3: 89.92% (96.58% after the second attempt), block 5: 89.43% (96.58% after the second attempt); and block 6: 81.79% (93.81% after the second attempt). A/not-A: Block 1 was completed in one attempt by 95.29%; block 2: 84.71% (96.30% after the second attempt); block 3: 80.59% (94.02% after the second attempt), block 5: 80.00% (93.53% after the second attempt); block 6: 70.59% (87.06% after the second and 95.88% after the third attempt).
4. Before receiving their result, participants indicated how much they tried to hide the card they chose and how much they engaged in faking during the test. Participants reported low scores for avoiding detection (Experiment 1: M = 2.83, SD = 2.12; Experiment 2: M = 3.30, SD = 2.40) and faking attempts (Experiment 1: M = 2.11, SD = 1.97; Experiment 2: M = 2.53,
SD = 2.28. For exploratory purposes, participants also indicated how important, negative, and positive they judged the card-labels, but these data are not reported upon (data on https://osf.io/swp27?view_only=c06b2668d5ff49a9a90d74bf3102333a26).

5. In preliminary analyses, block order showed no effect on the D score, F(1, 434) = 0.05, p = .831, f = 0.01, and was therefore not included in further analyses (see also Sartori et al., 2008).

6. We explored the role of testing native (native English speakers tested in English and native Dutch speakers tested in Dutch; n = 96, AUC = .92, 87–97) versus non-native speakers (non-native English speakers tested in English; n = 519, AUC = .88, .85–.91) by collapsing data for the A/B protocol of Experiments 1 and 2. The AUCs did not differ, D(151.87) = 1.18, p = .238.

Acknowledgement

We thank Tineke Slotegraaf for her assistance in piloting and setting up the web-aIAT.

Disclosure statement

No potential conflict of interest was reported by the authors.

References


