Behavior Research Methods

Detecting Lies about Consumer Attitudes using the Timed Antagonistic Response Alethiometer

Journal:	*Behavior Research Methods*
Manuscript ID:	BR-Org-13-127.R1
Manuscript Type:	Original Manuscript
Date Submitted by the Author:	n/a
Complete List of Authors:	Gregg, Aiden; University of Southampton, Psychology Mahadevan, Nikhila; University of Southampton, Psychology Edwards, Sonja; University of Southampton, Psychology Klymowsky, James; Seven Stones,

Running head: DETECTING LIES ABOUT ATTITUDES

Detecting Lies about Consumer Attitudes using the Timed Antagonistic Response Alethiometer

Aiden P. Gregg

Nikhila Mahadevan

Sonja E. Edwards

University of Southampton, UK

James Klymowsky

Seven Stones, UK

Author Note

Correspondence concerning this article should be addressed to Dr. Aiden P. Gregg, School of Psychology, University of Southampton, Southampton, SO17 1BJ. Email: aiden@soton.ac.uk

Word Count: text =8,251; abstract =119; references = 44

Abstract

The Timed Antagonistic Response Alethiometer (TARA) is a true-false statement classification task that diagnoses lying on the basis of slower average response speeds. Previous research (Gregg, 2007) found that a computer-based TARA was about 80% accurate when its statements conveyed demographic facts or religious views. Here, we tested the TARA's diagnostic potential when its statements conveyed *attitudes*—here, towards both branded and generic consumer products—across different versions of the TARA (Experiments 1a, 1b, and 1c), as well as across consecutive administrations (Experiment 2). Results generalized well across versions, and maximal accuracy rates exceeding 80% were obtained, although accuracy declined somewhat upon re-administration. Overall, the TARA shows promise as a comparatively cheap, convenient, and diagnostic index of lying about attitudes.

Keywords: TARA, lie detection, response speed, IAT, aIAT

Detecting Lies about Consumer Attitudes using

the Timed Antagonistic Response Alethiometer

People's intuitive capacity to detect lies is poor (Bond & DePaulo 2006, 2008). Hence, better technological alternatives have been sought. Chief among them is *polygraphy*, either in the form of the *Control Question Test* (CQT: Honts, Raskin, & Kircher, 2002), or the *Guilty Knowledge Test* (GKT: Lykken, 1998). Here, marked physiological arousal—driven by fabricating false answers or by cognizing correct ones—signals lying. However, both techniques have problems. For example, the CQT is arguably insufficiently specific (Moore, Petrie, & Braga, 2003), and the GKT, insufficiently sensitive (Vrij, 2008). More basically, both are relatively cumbersome and costly to administer, making them unsuitable for use in survey research.

Yet lying occurs in surveys too. One common cause is *social desirability bias* (Paulhus, 2002; Podsakoff, MacKenzie, Lee, & Podsakoff, 2003), when sensitive topics are being researched (Tourangeau & Ting, 2007). Even if social desirability is usually greater face-to-face (Christian, Dillman, & Smyth, 2008), lying still occurs online (e.g., about credit history; Karlan & Zinman, 2008), and may even be likelier online despite anonymity being assured (Naquina, Kurtzbergb, & Belkinc, 2010).

Hence, some alternative method of detecting lying on surveys—that delivers at least conventional levels of diagnostic accuracy while at the same time being both cheaper and handier to administer—would be welcome.

Response Speed as an Index of Deception

Whereas the CQT and GKT capitalize on physiological reactions to the recognized truth or falsity of assertions, it is also possible to capitalize on how long it takes to generate or process such assertions. That is, *response speed* may also diagnose lying. Recent research—going beyond conflicting prior meta-analyses

(DePaulo et al., 2003; Sporer & Schwandt, 2006)—now abundantly confirms that, when responding to direct inquiries in a structured manner, people take longer on average to lie than to tell the truth (Gregg, 2007; Holden, 1998; Sheridan & Flowers, 2010; Walczyk, Schwartz, Clifton, Adams, Wei, & Zha, 2005; Vendemia, Buzan, & Green, 2005). The underlying reason is cognitive: telling a lie, all else equal, requires more processing resources than telling the truth—resources devoted, amongst other things, to deciding to lie and to constructing a falsehood (Walczyk et al., 2005). Consistent with this account, imposing a cognitive load—say by having respondents recount a narrative backwards rather than forwards (Vrij et al., 2008)—permits liars to be better differentiated from truth-tellers (Vrij, Granhag, Mann, & Leal, 2011).

However, a workable lie detector must discriminate between lying and truth-telling, not only statistically, but also sensitively. Although the effect sizes yielded by simple response speed techniques may be conventionally large (e.g., Cohen's *d*s 39 to .90; see Gregg, 2007, p. 638), they often entail only modest diagnostic accuracy. Moreover, such techniques are susceptible to countermeasures (Robie et al., 2000), and may not detect lies that are well-rehearsed (DePaulo et al., 2003) or that concern mundane facts (Sporer & Schwandt, 2006). Furthermore, although more accurate GKTs incorporating response speed have been devised (e.g., Verschuere, Crombez, Degrootte, & Rosseel, 2010), the detection of concealed information is mainly of forensic interest. Accordingly, a more general-purpose lie detector based on response speed, and affording high levels of discrimination, is still desirable. Gregg (2007) sought to devise one: the *Timed Antagonistic Response Alethiometer*, or *TARA*.

The Timed Antagonistic Response Alethiometer (TARA)

The TARA works by making lying about some target topic harder than telling the truth about it. Everything else equal, lying must then take longer than telling the truth. The resulting difference in average response speed differentiates liars from truth-tellers. Importantly, the TARA does *not* merely capitalize upon the fact that spontaneous lying is harder than spontaneous truth-telling. Rather, it manufactures an *artificial* situation in which lying *becomes* harder than truth-telling.

TARA respondents complete two tasks simultaneously. For truth-tellers, these tasks are *compatible*. For liars, however, these tasks are *incompatible*. Hence, the two tasks comprising the TARA are harder for liars to complete than for truth-tellers. Because respondents must complete both tasks accurately, and accuracy is achieved at the expense of speed (Wickelgren, 1973), liars must proceed more slowly than truth-tellers, all else equal.

Procedurally, the TARA takes the form of a binary classification task. Respondents classify statements into the categories *True* and *False*. On computer, this entails pressing one of two keys whenever a statement appears. The key on the right corresponds to *True*, the key on the left to *False*. On-screen labels, on the same side as the relevant keys, reinforce these correspondences.

Respondents classify, on alternate trials, two different types of statement, *control* and *target*. The classification of each type of statement constitutes a separate task. In effect, then, respondents complete two tasks simultaneously. Depending on the classification strategy adopted, these tasks then end up being either compatible or incompatible.

Control statements are obviously true or false (e.g., *1 is a number* / *1 is a letter*). Barring accidental errors, no normal adult could plausibly misclassify them. Hence, truth-tellers and will classify all such statements identically. Control statements have nothing to do with the topic under investigation. In contrast, target statements pertain directly to that topic. Furthermore, truth-tellers and liars classify such statements differently. Consider the statements *I am X* and *I am not X*. The truth of either statement entails the falsity of the other. Hence, for every respondent, one statement will be true and the other false. However, truth-tellers and liars differ in which statements they *indicate* to be true or false. Among respondents who really are X, truth-tellers classify *I am X* as true and *I am not X* as false, whereas liars do the reverse. Similarly, among respondents who really are not X, truth-tellers classify *I am not X* as true and *I am X*, whereas again liars do the reverse. In both cases, the truth-tellers classify the true statement as true and the false statement as false, whereas liars classify the true statement as false and the false statement as true.[1]

Here, then, is the overall situation artificially manufactured by the TARA. Truth-tellers classify both control and target statements "the right way". In contrast, liars classify control statements "the right way" but target statements "the wrong way". Thus, on alternate trials, truth-tellers can adopt a consistent statement classification strategy—if true, indicate true; if false, indicate false. However, liars must adopt two inconsistent strategies, one for control statements—if true, indicate true; if false, indicate false—and another for target statements—if true, indicate false; if false, indicate true. The processing load imposed by having to switch strategy on alternate trials makes lying on the TARA more cognitively demanding than truth-telling. Given that respondents are instructed to classify the statements without making errors, the difference in cognitive demand results in a difference in average response speed, permitting liars and truth-tellers to be empirically distinguished.

Such is the theory. How well does the TARA work in practice? Gregg (2007) assessed its capacity to distinguish between the honest and dishonest reporting of personal details (e.g., age, gender). The TARA successfully did so (a) within the same group of participants (Study 1), and (b) between different groups of participants

(Study 2). In Study 1, *every* participant took longer when responding dishonestly than when responding honestly. In Study 2, levels of discriminative accuracy ranged from 79% to 100%, and Cohen's *d*s for the differences, from 2.47 to 2.89. Comparable values were obtained in a follow-up study concerning religious faith.

The Present Research: An "Attitudinal" Variant of the TARA

Gregg (2007) established that the TARA could distinguish between the honest or dishonest reporting of *beliefs* people had about themselves. Here, the task featured target statements linking an object (e.g., the self) with some *semantic attribute* (e.g., is male / female). But can the TARA equally distinguish between the honest or dishonest reporting of *evaluations* people have about attitude objects? Here, the task would feature target statements linking an object (e.g., a consumer product) with some *valenced attribute* (e.g., is positive / negative). If so, then the TARA might find a useful application in survey research, where opinions about issues are investigated as often as facts about self, and where honest responding cannot always be assured.

For example, in the consumer domain, the desire to create a good impression—which is already known to prompt unhealthy behaviors (Leary, Tchividjian, & Kraxberger, 1994)—might equally prompt deceptive reports, for example, false denials of liking tempting but unhealthy foods (e.g., hamburgers), or false affirmations of liking bland but healthy foods (e.g., broccoli). Equally, frank reporting of liking for products such illegal drugs or sexual aids is hardly guaranteed. Furthermore, correlations with standard measures of social desirability bias in market research do not always control for it, and attempts to do so may even reduce validity (Steenkamp, de Jong, & Baumgartner, 2010).

It bears mention that researchers have attempted to address lying about attitudes in two other noteworthy ways. First, they have attempted to facilitate the honest reporting of attitudes by falsely convincing respondents that they are hooked up to an infallible lie detector (i.e., via the "bogus pipeline" procedure; Jones & Sigall, 1971). Second, they have attempted to detect, by various indirect means, the underlying positive and negative valences that respondents associate with targets, thereby circumventing self-reports (i.e., via "implicit measures"; Wittenbrink & Schwarz, 2007). Both methods have met with partial success: the bogus pipeline reduces socially desirable responding (Roese & Jamieson, 1993) and implicit measures of attitude predict behavior above and beyond self-reports (Greenwald, Poehlman, Uhlmann, & Banaji, 2009). However, neither method diagnoses lying about attitudes *per se*. Moreover, the bogus pipeline is as unwieldy as a bona fide polygraph, and underlying associations need not correspond to conscious evaluations (Gawronski, 2009). Accordingly, an "attitudinal" TARA would still be a useful addition to a survey researcher's toolkit.

Experiment 1

In our first experiment, we tested whether and to what extent the TARA can detect whether participants are lying or telling the truth about the attitudes they hold. For simplicity and relevance, we choose a consumer product as an attitude object. In addition, we tested not just one "attitudinal" TARA, but three versions of it: one version conducted on computer (Experiment 1a); another making use of paper-and-pencil materials (Experiment 1b); and a third involving the sorting of playing cards (Experiment 1c). We reasoned that positive and comparable findings across all three versions would testify to the validity and robustness of the TARA's modus operandi, as well as to its flexibility in deployment.

Method

Consumer product. To assess the diagnostic accuracy of the proposed "attitudinal" TARA it was essential for us to establish the ground truth about participants' attitudes. Hence, we selected a consumer product towards which it could be *assumed* with near certainty that participants would reveal their true attitudes. We also selected a product that participants would not only be familiar with, but might tend either to strongly like or to strongly dislike, to test our hypothesis cleanly.

To satisfy both desiderata, we selected *Marmite*™. This is a classic British food spread, sold in small jars—a sticky, dark-brown paste possessing a distinctive flavour, both salty and savoury. No shame attaches to liking or disliking it: having either reaction is considered a matter of personal taste. Nonetheless, evaluations of *Marmite*™ are reputedly polarized: the product is marketed under the slogan "love it or hate it". As such, it was a convenient product with which to test the "attitudinal" TARA, likely to yield comparable numbers of lovers and haters.

Participants and recruitment. Participants in Experiment 1a comprised mainly undergraduate psychology students at the University of Southampton, UK, taking part for course credit. They completed a computer-based TARA, programmed in Authorware 7.0 (2000), and hosted online at *http://www.mindstudies.org/*. Full instructions were provided, and participants completed it at their own discretion.

Participants in Experiment 1b comprised volunteers recruited at a public library in Kenilworth, UK. They completed a paper-and-pencil TARA—a booklet containing ten A4-size pages—with a rollerball pen, leaning on a table, seated in a quiet corner, guided by the experimenter.

Participants in Experiment 1c mainly comprised undergraduate students at the University of Southampton, UK, who volunteered to take part. They completed the card-sorting task on a table in a quiet location on campus, guided by the experimenter.

The number of participants, and percentages of males, in Experiments 1a, 1b, and 1c, were respectively 135 (25%), 60 (42%), and 65 (45%). In addition, the means (and standard deviations) of participants' ages were respectively 22.8 (8.7), 23.8 (8.2), and 25.7 (9.8). Demographic information was collected in advance of completing the TARA.

Format, content, and design. The format, content, and design of Experiments 1a, 1b, and 1c were identical or near-identical.

First, after providing demographic information participants in all experiments indicated accurately (a) whether they liked or disliked *Marmite*™—by dichotomously indicating one or the other—and (b) how much they liked or disliked *Marmite*™–by rating their degree of liking for or dislike of it on a scale ranging from 1 (*I hate Marmite*) to 7 (*I love Marmite*).

Second, the TARAs in all three experiments featured the same two category pairs (*True*, *False*) and the same two sets of items (i.e., six target statements and six control statements: see Appendix, upper). The target statements were thematically homogeneous: the three items asserting liking for *Marmite*™ were semantically equivalent; the three items asserting dislike for *Marmite*™ were semantically equivalent; and the former meant the exact opposite of the latter.[2]

Third, each TARA consisted of 48 trials, such that items appeared in the same pseudo-random order for all participants (see Appendix, lower). This order satisfied four constraints simultaneously: (a) target statements always alternated with control statements; (b) all 12 basic statements appeared before any one reappeared; (c) no two identical statements appeared sequentially with fewer than three intervening items, and (d) no more than three statements belonging to the same category appeared sequentially (whether lying or telling the truth). The first constraint served to maximize response incompatibility when lying. Subsequent constraints served to prevent items from clumping together in a manner that might facilitate responding independently of levels of response incompatibility.

Fourth, all three experiments featured the same design. In each experiment participants completed practice trials, to familiar themselves with the procedure, and to ensure that they regarded both control and target statements as coherent units. They then completed the same TARA twice: once honestly, and once dishonestly. In each case, participants were told to proceed as quickly as they could without making errors, but to ignore any errors if they made them and to proceed to the next trial (trials advanced automatically in Experiment 1a). Order of honest and dishonest TARAs was counterbalanced across participants. In addition, the items to be classified appeared in "forwards" order the first time around (i.e., A-to-Z), but in "backwards" order the second time around (i.e., Z-to-A). This flipping around of item order prevented the pseudo-random sequence from being learned in the interim.

The key index for each TARA was mean response speed per trial. In Experiment 1a, the computer program automatically recorded response speeds for all 48 key presses. In Experiments 1b and 1c, the experimenter recorded by stopwatch how long it took participants to check all 48 boxes, or to sort all 48 cards, respectively. In addition, to avoid exclusion, participants also had to attain a satisfactory level of classification accuracy in both TARAs: 75% or greater.

Mean response speeds across different TARAs were then compared with a view to (a) establishing whether or not the new "attitudinal" TARA worked in principle—by testing for within-subject differences on the honest and dishonest TARAs; and (b) estimating how accurately the new "attitudinal" TARA might distinguish liars from truth-tellers in practice—by testing for between-subject differences on honest and dishonest TARAs. Note that the latter difference could be tested for twice, given that all participants completed two TARAs.

Procedure and materials.

In Experiment 1a, participants classified on-screen statements, located in the center of the display, by pressing keys on either side of a keyboard. In Experiment 1b, participants classified printed statements, arranged as a vertical list, by checking adjacent boxes on either side of a page. In Experiment 1c, participants classified statements on cards, located in a hand-held deck, into discard piles on either side of a table. (Figure 1 features photographs of the relevant materials.) In all cases, a rightward response indicated (whether honestly or dishonestly) that a statement was true, a leftward response that it was false. This correspondence was reinforced in each case by having permanent labels for *True* and *False* located above the statements to be classified (on the computer screen, booklet page, or table).

Finally, various procedural niceties served to optimize the measurement of response speed. For example, in Experiment 1a, before beginning each TARA, participants were reminded to place the fingers of either hand directly over the response keys, to facilitate rapid responding. In Experiments 1b and 1c, each TARA began with a verbal countdown, and ended with participants saying "done", to facilitate the accurate stopping and starting of the stopwatch.

Results

Data cleaning and reduction. In Experiment 1a, 26 participants (19%) were excluded for one or more of the following reasons: 5 for rating *Marmite*™ at the scale midpoint; 10 for being non-native English speakers; and 15 for completing a TARA more than once (having made too many errors). In Experiment 1b, 7 participants (12%) were excluded: 3 for rating *Marmite*™ at the scale midpoint; 1 for being a non-native English speaker; and 4 for making too many errors in either TARA. In Experiment 1c, 6 participants (9%) were excluded: 2 for rating *Marmite*™ at the scale midpoint; and 4 for dropping at least one card in either TARA.

Given the nature of materials, data were available at the level of individual trials Experiment 1a, but not in Experiments 1b and 1c. To maximize comparability across all three TARAs, response times from the TARA in Experiment 1a were only lightly processed. Obvious outliers on individual trials (< 350 ms or > 6000 ms) were replaced by the median response time across trials, but no nonlinear transformations were performed. Data from trials on which erroneous classifications were made were also retained. Mean response time to classify statements (i.e., total time taken / 48) served as the index of TARA performance in all experiments.

No response time penalties were applied in proportion to the number of classification errors made because the latter did not correlate significantly *negatively* with mean response time, either in Experiment 1a ($r_{HONEST}[107]= -.13$, $p = .17$; $r_{DISHONEST}[107] = .14$, $p = .14$), in Experiment 1b ($r_{HONEST}[51]= .15$, $p = .30$; $r_{DISHONEST}[51] = .18$, $p = .20$), or in Experiment 1c ($r_{HONEST}[57]= .21$, $p = .10$; $r_{DISHONEST}[57] = .29$, $p = .03$). If anything, the trend, at least in dishonest TARAs, suggests slower speeds and lesser accuracy covary positively rather than trade off.

Attitudes towards *Marmite*™. Consistent with popular lore, ratings of *Marmite*™ exhibited a bimodal distribution in all three experiments (see Figure 2, top). In none of the three experiments, however, did participants show any overall evaluative bias for or against *Marmite*™ (Experiment 1a: $t[108] = 1.25$, $p = .21$; Experiment 1b: $t[52] = -.92$, $p = .36$; Experiment 1c: $t[58] = -1.26$, $p = .21$).

Within-subject differences.

Overall effects. In all three experiments, the same set of participants took significantly (all *p*s < .0001) and substantially (all *d*s > 1.11) longer to complete the TARA when lying about their attitudes towards *Marmite*™ than when telling the truth about them (Table 1, upper). Thus, the "attitudinal" TARA worked in principle: regardless of the materials used to run it, the vast majority of participants exhibited the predicted effect. In terms of effect size, the paper-and-pencil version modestly exceeded the computer version, which modestly exceeded the card-sorting version.

Furthermore, participants made slightly more errors in the dishonest TARA than in the honest TARA, consistent with slower speeds and lower accuracy both reflecting processing difficulty (Experiment 1a: respective *M*s = 3.50, 1.74, $t[108]= 5.70$, $p < .0005$; Experiment 1b: respective *M*s = .87, .25, $t[52]= 3.40$, $p < .001$; respective *M*s = 2.31, .73, $t[58]= 5.95$, $p < .0005$). Nonetheless, accuracy rates were very high across the board, even in dishonest TARAs (i.e., 93%, 98%, and 95% across Experiments 1a, 1b, and 1c). This has implications for the nature of the mechanism underlying TARA effects. It means that the overall response speed effects, reported above, cannot plausibly be ascribed, wholly or mainly, to greater cognitive confusion regarding how to respond in dishonest TARAs than in honest TARAs. More plausibly, participants understood equally well what to do in dishonest TARAs, but found it objectively more difficult to do it, and accordingly went more slowly. Their occasional errors can be better put down to their complying only imperfectly with the instruction to avoid errors. Backing up this interpretation, even among those participants who made *zero* errors—and so who could not conceivably be described as confused—significant within-subject differences between TARAs emerged, despite such participants being far fewer in number (Experiment 1a: $M_{DIFF} = 464$ ms, $t[4]= 2.68$, $p = .055$; Experiment 1b: $M_{DIFF} = 495$ ms, $t[23]= 8.06$, $p < .0005$; Experiment

1c: M_{DIFF} = 576 ms, $t[10]$= 3.82, $p < .005$). Moreover, participants who did, or did not, make errors, showed equivalent within-subject differences between TARAs (Experiment 1a: $t[107]$= -.42, $p = .67$; Experiment 1b: $t[51]$= .33, $p = .75$; Experiment 1c: $t[57]$= .72, $p = .47$).

Moderation tests. We tested whether gender (coded female = 0, male = 1), age (in years), attitude towards target (coded 1 to 7), and the extremity of that attitude (coded 1 to 3) moderated average response speed for both the honest and dishonest TARAs, as well as for the within-subject difference between them (Table 2, upper).

No variable significantly moderated the within-subject difference. Greater age predicted numerically slower responses on all indices, which attained significance for both the honest and dishonest computer-based TARAs (as in Gregg, 2007), and for the dishonest card-sorting TARA. In addition, males, and dislikers of *Marmite*™, completed the honest computer-based TARA more slowly; but gender and attitude towards *Marmite*™ predicted no other effects. Interestingly, extremity of attitudes towards the *Marmite*™ did *not* significantly predict performance on *any* TARA. This suggests that, at least for *Marmite*™, whether one tells the truth or tells a lie may matter more than the magnitude of the truth or lie one tells.

Between-subject differences.

Overall effects. In all three experiments, as predicted, sets of participants who lied about their attitudes towards *Marmite*™ took longer to complete the TARA than matched sets participants who told the truth about them (Table 3, upper). The difference was significant and substantial both for the TARA completed first (all ps < .0005, all ds > 1.57) and for the TARA completed second (all ps < .001, all ds > 1.00). Thus, the “attitudinal” TARA showed promise as a tool to be used in practice: regardless of the materials used to run it, sizeable diagnostic differences emerged between liars and truth-tellers. In terms of effect size, the paper-and-pencil and computer versions modestly exceeded the card-sorting version.

Percentage accuracy. The conventional way to quantify the diagnosticity of a lie detector is in terms of its percentage accuracy when discriminating liars from truth-tellers. This overall value is typically broken down further into four component values: true positives (i.e., percentage of liars correctly marked as liars: "hits"), false positives (i.e., percentage of truth-tellers mistakenly marked as liars: "false alarms"), true negatives (i.e., percentage of truth-tellers correctly marked as truth-tellers: "correct rejections"), and false negatives (i.e., percentage of liars mistakenly marked as truth-tellers: "misses"). In the absence of a separate empirical investigation—where diagnosticity would be directly assessed in a given sample against a preset criterion—percentage accuracy and its component values can be estimated by simulation from existing data.

The estimation method we chose was a straightforward leave-one-out simulation. On each iteration, we randomly selected, with replacement, one participant as the target whose status as a liar or truth-teller was to be inferred. The remainder of the sample was then divided into two groups: liars and truth-tellers. We calculated the grand mean of the mean response times for both. The mean of these two grand means—a value falling roughly midway between the distributions for liars and truth-tellers—served as the criterion for inferring the status of the target. Specifically, if the target were a liar, the result was defined as a true positive if his mean response time lay above the criterion, and as a miss if it lay below it; if the target were a truth-teller, the result was defined as a true negative if his mean response time lay below the criterion, and as a false positive if it lay above it. We tallied results across 30,000 iterations, and expressed them as percentages. Overall percentage accuracy was calculated as the mean of the percentage of true positive and true negatives. This method of estimating discrimination can be regarded as assessing the accuracy of the TARA when applied to any new individual drawn from exactly the same population as that from which TARA performance norms are previously derived.

Maximal values achieved were high, with overall accuracy exceeding 80%, while minimal values achieved were respectable, with overall accuracy around 70% (Table 4, upper). In addition, all TARAs showed a bias towards correctly identifying truth-tellers (i.e., specificity) at the expense of correctly identifying liars (i.e., sensitivity)—arguably a desirable bias for any lie detector to exhibit. Finally, echoing earlier findings, the computer-based and paper-and-pencil versions achieved somewhat higher levels discrimination than the card-sorting version on the TARA completed first. However, roughly equivalent levels of discrimination were achieved across all versions on the TARA completed second.

Discussion

In Experiment 1a, 1b, and 1c, we tested whether, and to what extent, three different versions of the TARA—a lie detector designed to diagnose deception on the basis of slower responses Gregg (2007)—could distinguish between the honest and dishonest reporting of attitudes towards a consumer product, namely, *Marmite*™. Regardless of which version of the TARA was used—computed-based, paper-and-pencil, or card-sorting—two clear-cut findings emerged: (a) the same set of participants went markedly more slowly when lying about their attitudes than when telling the truth about them; and (b) sets of participants who lied about their attitudes went markedly more slowly than sets of participants who told the truth about them. Thus, across varying formats, the TARA methodology consistently showed promise as an index of falsely expressed attitudes. Moreover, its results were neither strongly nor consistently moderated by gender, age, attitude towards *Marmite*™, or extremity of attitude towards *Marmite*™. Moreover, simulation tests estimated its potential accuracy rates to match those of leading lie detectors, such as the CQT and GKT (Vrij, 2008).

Experiment 2

Experiments 1a, 1b, and 1c showed that the TARA methodology worked across three different versions. However, they featured only a single attitude object, namely, *Marmite*™. The possibility therefore arises that the effects observed were exclusive or particular to that attitude object.

As mentioned above, *Marmite*™ is reputedly distinctive in that people are alleged by advertisers to have extremely bimodal attitudes towards it. Indeed, the patterns of attitudes towards *Marmite*™ we observed did take broadly bimodal form (Figure 2, upper). Might the TARA therefore diagnose lying and truth-telling only with respect to objects like *Marmite*™? We address the question empirically below. But before doing so, we make several pertinent points.

First, although attitudes towards *Marmite*™ were bimodally distributed in our samples, they were hardly devoid of variance. Our samples featured reasonable numbers of participants holding mild, moderate, and extreme attitudes towards *Marmite*™, both favorable and unfavorable.

Second, we took the trouble to investigate empirically in Experiments 1a, 1b, and 1c whether such attitudinal variance—specifically, the extremity of attitudes towards *Marmite*™–moderated performance on the TARA. In no case did it do so, neither when participants responded honestly, nor when they responded dishonestly. In other words, participants who held milder attitudes towards *Marmite*™ exhibited effects on the TARA equivalent to participants who held more extreme attitudes towards *Marmite*™. Thus, no evidence emerged to support concerns that the TARA only works when attitudes are extreme, when it could readily have emerged in three separate experiments.

Third—and more pragmatically—if people lack any attitude towards an object—that is, if they are wholly indifferent to it, then there is no attitude for them either to tell the truth or to lie about. Accordingly, the TARA would have no application in such cases. Hence, that it should not operate in such cases can hardly be a criticism of it. The only concern would be this: if, in a given sample, many members were indifferent to an attitude object, then many bogus TARA results might contaminate that sample, and thus impair its diagnosticity. In mitigation, however, it should be pointed out that even if indifference were the modal attitude, which is hardly guaranteed, such a modal attitude would still be but a fraction of the total distribution of attitude, except in very leptokurtic distributions. At any rate, there are no a priori grounds to believe that indifference should systematically bias the TARA towards signaling honesty or dishonesty, but only that it would add random error.

Nonetheless, we judged it prudent to test whether the TARA could diagnose honesty and dishonesty with respect to another attitudinal object, one that—although still being unlikely to arouse social desirability bias so that ground truth could be presumed—was liable to yield a less bimodal distribution of liking and disliking, and which was also better known. For this purpose, we chose a popular everyday product, still consumable but more generic: *coffee*.

In addition, we administered both TARAs, honest and dishonest, not once, but *twice* to participants in Experiment 2, in the same counterbalanced order. The purpose of doing so was to assess the impact on TARA performance of immediate prior experience with the TARA. We wondered: would the opportunity to familiarize oneself with, and gain practice at, lying and telling the truth on the TARA, reduce or eliminate any subsequent TARA effect (assuming that we could replicate the initial TARA effect)? Thus, the study tested not only the generality of the TARA across topic, but also its persistence over time. Furthermore, to the extent that TARA effects did persist, the design offered a means of estimating split-half reliability.

Method

Procedure and materials. For Experiment 2, we opted to use the paper-and-pencil version of the TARA. Accordingly, the procedure was identical to that reported in Experiment 1b, except that five different experimenters rather than the same one ran the study (to expedite its completion). The only difference in the materials was that (a) the word "coffee" was substituted for the word "*Marmite*" in the experimental booklets, and (b) two additional pages were added to the end of the booklet, to add a second pair of honest and dishonest TARAs (i.e., Pair II), to the first pair (i.e., Pair I). To prevent item-specific practice, and make items appears random—but also maximize the comparability of the two pairs—the vertical order in which 48 items appeared was inverted for Pair II relative to Pair I.

Participants. All 73 participants were volunteers whom the experimenters approached personally to take part. Most participants were undergraduate students from the University of Southampton campus but several were older ($M_{\text{AGE}} = 30.2$; $SD_{\text{AGE}} = 12.6$). The sample was roughly balanced for gender (44% male).

Results

Data cleaning and reduction. In Experiment 2, 13 participants (18%) were excluded for one or more of the following reasons: 7 for rating coffee at the scale midpoint; 1 for being a non-native English speaker; 6 for making too many errors in any TARA; and 2 for failing to complete the experimental booklet. Again, mean response time to classify statements (i.e., total time taken / 48) served as the index of TARA performance. As before, no penalties for classification errors were applied, as they did not correlate significantly negatively with mean response time, either in Experiment 2_{I} ($r_{HONEST}[58] = .11$, $p = .40$; $r_{DISHONEST}[58] = .49$, $p < .0005$), in Experiment 2_{II} ($r_{HONEST}[58] = .22$, $p = .10$; $r_{DISHONEST}[58] = .42$, $p < .001$).

Attitudes towards coffee. As expected, ratings of coffee exhibited a more normal distribution than ratings of *Marmite*™ (see Figure 2, bottom). In particular, the modal value (+1) lay near to the center of the distribution, and, unlike in Experiments 1a, 1b, and 1c, the two most extreme values (+3, -3) were less frequent than the next two most extreme values (+2, -2). Recalling earlier findings for *Marmite*™, participants showed no overall evaluative bias for or against coffee, $t(59) = 1.12$, $p = .27$.

Within-subject differences.

Overall effects. As predicted, the same set of participants took significantly longer to complete the TARA when lying about their attitudes towards coffee than when telling the truth about them (both *p*s < .0005; see Table 1, lower), both for Pair I, Pair II, and the average of the two. Thus, the "attitudinal" TARA also worked with respect to a new and more familiar consumer product that elicited a more normal distribution of liking and disliking. Accuracy rates were comparable to those in Experiments 1a, 1b, and 1c. Effect sizes ranged between being about half the size of, to being about one-third the size of, those obtained in Experiment 1b.

In addition, the within-subject difference was significantly lower for Pair II than for Pair I, $t(59) = 4.28$, $p < .001$, with the significant reduction for successive dishonest TARAs, $t(59) = 6.06$, $p < .0001$, being about twice the size of that for successive honest TARAs, $t(59) = 3.43$, $p < .001$. This suggests that prior experience with the TARA, whether lying or telling the truth, may somewhat reduce its diagnosticity. Nonetheless, even if aggregate performance shifted across consecutive TARAs, individual performance exhibited substantial consistency: honest TARAs, dishonest TARAs, and the within-subject difference between them, yielded large split-half reliabilities of $r = .79$, $r = .90$, and $r = .77$ respectively.[3] Such consistency attests to the intraindividual robustness of TARA effects.

Moderation tests. Once again, we tested whether gender, age, attitude towards target, and extremity of attitude moderated average response speed for both the honest and dishonest TARAs, as well as for the within-subject difference between them (Table 2, lower). Overall, more evidence of significant moderation emerged here than emerged in Experiments 1a 1b, and 1c. Greater age predicted slower responses on all indices. Unexpectedly, gender also emerged as a moderator: being male predicted significantly slower responses on all indices. Some inconsistent signs also emerged that attitude towards target, and the extremity of that attitude, significantly moderated TARA performance. The most consistent finding here, however, was that more extreme attitudes predicted *slower* performance on the honest TARA.

Between-subject differences.

Overall effects. Sets of participants who lied about their attitudes towards coffee took longer to complete the TARA than matched sets of participants who told the truth about them (Table 3, lower). For the TARA completed first, whether in Pair I and Pair II, the between-subject difference was significant and substantial; for the TARA completed second, the corresponding between-subject differences were still significant, but less substantial. Averaging across both pairs, the between-subject differences persisted. Again, effect sizes ranged between being about half the size of, to being about one-third the size of, those obtained in Experiment 1b.

Percentage accuracy. As in Experiments 1a, 1b, and 1c, the between-subject diagnosticity of the TARA was expressed—by means of an odd-one-out statistical simulation—in terms of true positives (identified liars), false positives (misidentified truth-tellers), true negatives (identified truth-tellers), and false negatives (misidentified liars). For the TARA completed first, Pair I values approached the high levels seen in Study 1b. Pair II values for the TARA completed first, and Pair I values for the TARA completed second, were somewhat reduced. Pair II values for the TARA completed second exceeded chance only modestly. Averaging across both Pair I and Pair II, intermediate values were obtained, again about one-half to two-thirds of those obtained in Experiment 1b (starting the count at chance: 50% accurate). In addition, the TARA showed the same bias towards specificity at the expense of sensitivity.

Discussion

In Experiment 2, we tested whether, and to what extent, the TARA could distinguish between the honest and dishonest reporting of attitudes towards a new consumer product—namely, coffee—lest the results of Experiments 1a, 1b, and 1c be specific to the use of *Marmite*™. Overall, Experiment 2 replicated earlier effects, both within-subject and between-subject, testifying to its generality. However, both the magnitude of the effects obtained and the diagnosticity of the TARA were affected by whether the TARA was being completed for the first or for the second time: performance levels approached those of Experiment 1b in the former case but fell behind them in the latter. Thus, prior experience with the TARA attenuated effects. In addition, Experiment 2 yielded some indications that TARA performance could be moderated by extraneous variables. However, apart from age—known to correlate inversely with reaction time—the effects observed were inconsistent (attitude towards target), lay in an unexpected direction (extremity of attitude), or were likely due to chance (gender).

General Discussion

Overall, the "attitudinal" TARA tested here showed promise as a lie detection tool for use in survey research. Its effects generalized either well or reasonably well across high-tech and low-tech versions, different attitude objects, and repeated administrations.

We now devote the remainder of this paper to discussing (a) variations in effect magnitude observed, and (b) similarities and differences between the TARA and another recently developed response speed task designed to assess deception.

Variations in Effect Magnitude

Four sources of observed variation in the magnitude of TARA effects merit comment: the TARA version used; the attitude object addressed; the demographic characteristics of the respondents; and prior experience with the TARA. We deal with these below, roughly in turn, but sometimes in tandem.

In Experiment 1, depending on the index analyzed, the computer-based and paper-and-pencil TARAs either yielded comparable effects (i.e., for the between-subject comparison), or one slightly outdid the other (i.e., for within-subject effects and accuracy estimates). The card-sorting TARA trailed behind somewhat in terms of maximal effects obtained. Yet at the same time, the card-sorting TARA yielded more robust effects for TARAs administered second. Thus, regardless of how the logic underlying the TARA was instantiated, roughly comparable effects emerged across the board.

The magnitude of effects observed in Experiment 2, despite all being statistically significant, ranged between one-half and two-thirds of those observed in Experiment 1a. In both experiments, the paper-and-pencil TARA was administered; hence, some other factor or factors must explain the disparity. One explanation is that the "attitudinal" TARA—although it works for attitude objects whose evaluations are distributed bimodally (i.e., *Marmite*™) or more normally (i.e., coffee)—nonetheless works somewhat better in the former case. An alternative explanation, however, is that the samples for Experiment 1b and Experiment 2 accidentally differed in some relevant way.

Age emerged as a moderator of TARA performance, either some the time (Experiment 1) or all of the time (Experiment 2), such that older participants took longer to complete the TARA. This finding echoes others from large-scale studies on choice reaction times (e.g., Der & Deary, 2006) and implies that age is liable to be a demographic confound worth controlling for (cf. Greewald, Nosek, & Banaji 2003). Furthermore, older participants showed a larger within-subject difference in Experiment 2 (but not in Experiment 1). This being the case—and given that participants in Experiment 2 were on average 10 years older than those in Experiment 1 (especially, Experiment 1b)—one might have expected, if anything, that TARA effects would have been larger in Experiment 2. However, they were smaller. Hence, age is unlikely to account for the difference in effect size between the experiments.

Experiment 2 (but not Experiment 1) also saw an unexpected moderator emerge: gender. Contrary to the findings from large-scale research (Der & Deary, 2006), males completed the TARA more slowly than females. This suggests that the participants in Experiment 2 may have been atypical. If so, their atypicality might also explain the lower magnitude of effects in Experiment 2.

In neither Experiment 1 nor Experiment 2 did the direction of participants' attitudes towards *Marmite*™ consistently moderate the size of effects observed. This is a reassuring result for an "attitudinal" TARA. However, Experiment 2 (but not Experiment 1) did find some evidence that having more moderate or extreme attitudes moderated the size of effects observed. Curiously, however, the evidence was at least as strong as that having more extreme attitudes slowed down the completion of the *honest* TARA as of the dishonest TARA. One might have imagined finding the opposite pattern—that stronger attitudes might have been "truer" and so have expedited completion of the honest TARA while retarding completion of the dishonest TARA. At all events, the data from neither Experiment 1 nor Experiment 2 offer any support for this pattern. Overall, then, the "attitudinal" TARA seems to detect "little" lies as well as "big" ones.

In Experiment 2, the diagnosticity of the TARA pair administered first (i.e., Pair I) markedly exceeded that of the TARA pair administered second (i.e., Pair II). This would appear to be a liability: experiential acquaintance with the TARA may be sufficient to undermine its capacity to distinguish robustly between liars and truth-tellers. Furthermore, determined practice, including on specific TARAs, might undermine that diagnosticity further.

Note, however, that this liability need not be realized. In practice, the "attitudinal" TARA is likely to be administered, often if not nearly always, to respondents who are unfamiliar both with the specific TARA being administered, and with the TARA methodology itself. It will not be used in forensic settings to decide the guilt or innocence of criminal suspects, but rather in research settings to estimate the veracity or mendacity of survey respondents.

Comparing and Contrasting the TARA and the aIAT

The issue of guilt or innocence brings us to the *Autobiographical IAT* (*aIAT;* Sartori, Agosta, Zogmaister, Ferrara, & Castiello, 2008)—another lie detector relying on response speed. It is instructive to compare and contrast the aIAT and TARA. To do so, it is useful to relate both to their common "ancestor": the *Implicit Association Test* (*IAT*; Carney, Nosek, Greenwald, & Banaji, 2007; Greenwald et al., 1998), designed to be an indirect measure of automatic association between category pairs.

The IAT works as follows. Respondents classify items—either words or pictures—into four categories (*A*, *B*, *X*, and *Y*). They do so by making either of two responses (e.g., *left* and *right*) as quickly as they can without making errors. In two critical blocks, the assignment rules for categories and responses are switched around. Specifically, if respondents go more quickly in one block (e.g., where [$A \mid X \rightarrow$ *left*] & [$B \mid Y \rightarrow$ *right*]) than in another (e.g., where [$A \mid Y \rightarrow$ *left*] & [$B \mid X \rightarrow$ *right*]), then they are deemed to automatically associate category pairs in one way (i.e., [A ↔ X] & [B ↔Y]) rather than another (i.e., [A ↔ Y] & [B ↔ X]).

The aIAT differs from the IAT, but resembles the TARA, in that respondents classify statements, not words or pictures. The aIAT also differs from the IAT, but again resembles the TARA, in that respondents classify some statements into the categories *True* and *False* (specifically, the *control* statements).

However, the aIAT and the IAT both differ from the TARA in that, whereas they both feature two critical blocks of trials to be compared within-subject, the TARA features only a single critical block of trials to be compared between-subject. In addition, the aIAT and the IAT both differ from the TARA in that, whereas four categories appear in both their critical blocks, only two categories appear in the TARA critical block (i.e., *True* and *False*).

More specifically: in the critical blocks of the aIAT, respondents classify control statements into the categories *True* or *False* but target statements into the categories *Innocent* or *Guilty*. In its first critical block, the assignment is [*True* | *Guilty* → *left*] & [*False* | *Innocent* → *right*]; in its second critical block, the assignment is [*True* | *Innocent* → *left*] & [*False* | *Guilty* → *right*]. Thus, respondents classify target statements on the aIAT in two ways: first, as if they have something to hide. Conceptually, therefore, the second critical block loosely corresponds to the TARA critical block. On the TARA, all respondents, whether actually responding honestly (i.e., "innocent" respondents) or dishonestly ("guilty" respondents), classify target statements as if they have nothing to hide.

The crucial difference is this: whereas all statements on the TARA are classified in terms of their *truth-value*, half the statements on the aIAT—that is, the target statements—are classified in terms of their *categorical meaning* (i.e., just as *all* stimuli on the original IAT are). For example, "I stole the jar of Marmite" would be classified under the category "guilty" whereas "I bought the jar of Marmite" would be classified under the category "innocent". This reflect the intended modus operandi of the aIAT: it was designed to reflect levels of association between the category pairs *True* and *False* and the category pairs *Guilty* or *Innocent*—levels of association assumed to depend in turn on the guilt or innocence of respondents. In contrast, the TARA does *not* rely upon any associations between categories. Instead, it relies—as outlined in the introduction section—only on the differential compatibility of classification strategies that respondents adopt by virtue of honestly or dishonestly classifying target statements in conjunction with accurately classifying control statements.

How then do the TARA and aIAT compare as potentially cheap, convenient, and diagnostic indices of deception, suitable for use in survey research?

In terms of cheapness, both techniques are indistinguishable, being equivalently low-tech. There is no reason to suppose that paper-and-pencil or card-sorting versions of the aIAT could not also be readily realized.

In terms of convenience, the TARA exhibits two relevant advantages, one arguably desirable and the other arguably decisive. First, the TARA features only *one* block of trials rather than two. Hence, it can be administered slightly more quickly and efficiently than the aIAT can. Second, target statements on the TARA are classified in terms of truth versus falsity rather than guilt versus innocence. Hence, the TARA is far more generally applicable than the aIAT. Indeed, the concepts of guilt or innocence simply *do not apply* in survey research, as no a priori value judgment is made about what respondents report, nor are respondents subjects in a criminal investigation. For example, how would one even code guilt or innocence with respect to attitudes expressed towards *Marmite*™ or coffee? Thus, the aIAT is primarily suited to forensic use, like recent response speed versions of the GKT (e.g., Verschuere et al., 2010). The TARA, in contrast, can take more general form, including as an "attitudinal" lie detector.

Finally, in terms of diagnosticity, both the TARA and aIAT have relative advantages and disadvantages. In principle, because the aIAT (like the original IAT) relies on the within-subject comparison of response speed, it can partly control for idiosyncratic variations in response speed that the TARA cannot (Greenwald et al., 2003). Second, whereas TARA effects are driven only by a single source of incompatibility, present in a single block, aIAT could be driven by two sources of incompatibility, each present in either of its pair of blocks. For example, relative to an innocent participant, a guilty participant may proceed both (a) more quickly on the block where [*True* | *Guilty* → *left*] & [*False* | *Innocent* → *right*] and (b) more slowly on the block where [*True* | *Innocent* → *left*] & [*False* | *Guilty* → *right*]. The interactive pattern of results obtained for some studies (Sartori et al., 2008) suggests that this is the case. Thus, aIATs may yield comparatively stronger effects.

Yet the aIAT's potential diagnostic advantages may not always be realized. To understand why, consider what is required to "cheat" the TARA and the aIAT. To cheat on the TARA, dishonest respondents must go more quickly; and to do that, they must overcome the incompatibility imposed by the task. To cheat on the aIAT, however, guilty respondents can adopt either of two tactics: they can go more quickly in the block where [*True* | *Innocent* → *left*] & [*False* | *Guilty* → *right*], by again overcoming the incompatibility imposed; or they can go more slowly in the block where [*True* | *Guilty* → *left*] & [*False* | *Innocent* → *right*], simply by intentionally slowing down. Adopting the latter tactic, unavailable on the TARA, is easier than adopting the former, as slowing down poses no special challenge (cf. Fiedler & Bluemke, 2005). Hence, it is *in principle* easier to cheat the aIAT than the TARA. Moreover, respondents given simple instructions can *in practice* cheat the aIAT (Verschuere, Prati, & De Houwer, 2009). Still, respondents they might not spontaneously figure out how to do so in the absence of such instructions (cf. Kim, 2003). Furthermore, algorithms have recently developed that can distinguish well between faked and frank responding on the aIAT (Agosta, Ghirardi, Zogmaister, Castiello, & Sartori, 2011; see also Cvencek, Greenwald, Brown, Gray, & Snowden, 2010), so this may decrease the aIAT's greater structural susceptibility to fakery.

References

Agosta, S., Ghirardi, V., Zogmaister, C., Castiello, U., & Sartori, G. (2011). Detecting fakers of the autobiographical IAT. *Applied Cognitive Psychology*, *25*, 299–306..

Authorware (Version 7.0) [Computer software]. (2000). San Jose, CA: Adobe Macromedia LLC.

Bond, C. F. Jr., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review*, *10*, 214-234.

Bond, C. F. Jr., & DePaulo, B. M. (2008). Individual differences in detecting deception: Accuracy and bias. *Psychological Bulletin*, *134*, 477-492.

Carney, D. R., Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). Implicit Association Test (IAT). In R. Baumeister & K. Vohs (Eds.), *Encyclopedia of Social Psychology* (pp. 463-464). Thousand Oaks, CA: SAGE

Christian, L., Dillman, D., & Smyth, J. (2008). The effects of mode and format on answers to scalar questions in telephone and web surveys. In J. M. Lepkowski et al. (Eds.), *Advances in telephone survey methodology* (pp. 250-275). Hoboken, NJ: Wiley-Interscience.

Cvencek, D., Greenwald, A. G., Brown, A. S., Gray, N. S., & Snowden, R. J. (2010). Faking of the Implicit Association Test is statistically detectable and partly correctable. *Basic and Applied Social Psychology*, *32*, 302-314.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin*, *1*, 74-118.

Der, G., & Deary, I. J. (2006). Age and sex differences in reaction time in adulthood: Results from the United Kingdom health and lifestyle survey. *Psychology and Aging*, *21*, 62-73.

Fiedler, K., & Bluemke, M. (2005). Faking the IAT: Aided and unaided response control on the implicit association tests. *Basic and Applied Social Psychology*, *27*, 307–316.

Gawronski, B. (2009). Ten frequently asked questions about implicit measures and their frequently supposed, but not entirely correct answers. *Canadian Psychology*, *50*, 141-150.

Gregg, A. P. (2007). When vying reveals lying: The Timed Antagonistic Response Alethiometer. *Applied Cognitive Psychology*, *21*, 621–647.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. K. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74,* 1464-1480.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*, 197-216.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, *97*, 17–41.

Holden, R. R. (1998). Detecting fakers on a personnel test: Response latencies versus a standard validity scale. *Journal of Social Behavior and Personality*, *13*, 387–398.

Honts, C. R., Raskin, D. C., & Kircher, J. C. (2002). The scientific status of research on polygraph techniques: The case for polygraph tests. In D. L. Faigman, D. H. Kaye, M. J. Saks, & J. Sanders (Eds.), *Modern scientific evidence: The law and science of expert testimony* (Vol. 2, pp. 483–538). St. Paul, MN: West Law.

Jones, E. E., & Sigall, H. (1971). The bogus pipeline: A new paradigm for measuring affect and attitude. *Psychological Bulletin*, *76*, 349–364.

Karlan, D., & Zinman, J. (2008). Lying about borrowing. *Journal of the European Economic Association*, *6*, 510-521.

Kim, D. Y. (2003). Voluntary controllability of the Implicit Association Test (IAT). *Social Psychology Quarterly*, *66*, 83-96.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*, 213-236.

Leary, M. R., Tchividijian, L. R., & Kraxberger, B. E. (1994). Self-presentation can be hazardous to your health: Impression management and health risk. *Health Psychology*, *13*, 461-470.

Lykken, D. (1998). *A tremor in the blood: The uses and abuses of the lie detector*. Reading, MA: Perseus books.

Meston, C. M., Heiman, J. R., Trapnell, P. D., & Paulhus, D. L. (1998). Socially desirable responding and sexuality reports. *Journal of Sex Research*, *35*, 148-157.

Moore, M. H., Petrie, C. V., & Braga, A. A. (2003). *The polygraph and lie detection*. Washington DC: The National Academies Press.

Naquina, C., Kurtzbergb, T., & Belkinc, L. (2010). The finer points of lying online: E-mail versus pen and paper. *Journal of Applied Psychology*, *95*, 387-394.

Podsakoff, P. M., MacKenzie, S. B., Lee, J., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*, 879-903.

Paulhus, D. L. (2002). Socially desirable responding: the evolution of a construct. In H. I Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in* *psychological and educational measurement* (pp.46-69). Mahwah NJ: Lawrence Erlbaum.

Robie, C., Curtin, P. J., Foster, T. C., Phillips, H. L., Zbylut, M., & Tetrick, L. E. (2000). The effects of coaching on the utility of response latencies in detecting fakers on a personality measure. *Canadian Journal of Behavioral Science*, *32*, 226–233.

Roese, N., & Jamieson, D. (1993). Twenty years of bogus pipeline research: A critical review and meta-analysis. *Psychological Bulletin*, *114*, 363-375.

Sartori, G., Agosta, S., Zogmaister, C., Ferrara, S. D., & Castiello, U. (2008). How to accurately detect autobiographical events. *Psychological Science*, *19*, 772-780.

Sheridan, M. R., & Flowers, K. A. (2010). Reaction times and deception - the lying constant. *International Journal of Psychological Studies*, *2*, 26-40.

Sporer, S. L., & Schwandt, B. (2006). Paraverbal indicators of deception: A meta-analytic synthesis. *Applied Cognitive Psychology, 20*, 421–446.

Steenkamp, J. E. M., de Jong, M. G., & Baumgartner, H. (2010). Socially desirable response tendencies in survey research. *Journal of Marketing Research*, *47*, 199-214.

Tourangeau, R., & Ting, Y. (2007). Sensitive questions in surveys. *Psychological Bulletin*, *133*, 859-883.

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

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

Verschuere, B., Prati, V., & De Houwer, J. (2009). Cheating the lie detector: Faking the autobiographical IAT. *Psychological Science*, *20*, 410-413.

Vrij, A. (2008). *Detecting lies and deceit: Pitfalls and opportunities*. Chichester: Wiley.

Vrij, A., Granhag, P. A., Mann, S., & Leal, S. (2011). Outsmarting the liars: Towards a cognitive lie detection approach. *Current Directions in Psychological Science*, *20*, 28-32.

Vrij, A., Mann, S., Fisher, R., Leal, S., Milne, B., & Bull, R. (2008). Increasing cognitive load to facilitate lie detection: The benefit of recalling an event in reverse order. *Law and Human Behavior*, *32*, 253–265.

Walczyk, J. J., Schwartz, J. P., Clifton, R., Adams, B., Wei, M., & Zha, P. (2005). Lying person-to-person about life events: A cognitive framework for lie detection. *Personnel Psychology*, *59*, 141–170.

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica*, *41*, 67–85.

Wittenbrink, B. & Schwarz N. (Eds.) (2007). *Implicit measures of attitudes*. New York: Guilford Press.

Footnotes

[1] It should perhaps be clarified that all target statements of one sort [e.g., *I am male*; *I am a man*; *I am a guy*], and all target statements of another sort [e.g., *I am female*; *I am a woman*; *I am a lady*]—are designed, in each case, to be classified *identically*. That is, target statements of one type are *all* classified either honestly or dishonestly, as a coherent unit, and target statements of another type are *all* classified either dishonestly or honestly, as a coherent unit. The detection of truth-telling or deception occurs *across* sets of items, not *between* individual items.

[2] We established this empirically. A separate group of participants ($N = 91$) rated each of the six target statements for how much they agreed or disagreed with them (1 = *Completely Disagree*, 7 = *Completely Agree*). After reverse-scoring ratings for the three items asserting dislike, the degree of internal consistency across all six ratings was estimated. The near maximal value obtained (α =.97) implied that the target statements were thematically homogeneous, as intended.

[3] Experiment 1a, whose data consisted of multiple individual response times, also afforded the possibility of estimating the internal consistency of the TARA blocks. Split-half reliability—the correlation between the average RTs across the first 24 response times and the average RT across the last 24 response times—was respectively $r = .76$, $r = .69$, and $r = .52$—for the honest TARA block, the dishonest TARA blocks, and the within-subject difference between them, respectively. If the split-halves were computed across adjacent pairs of trials (e.g., 1,2,5,6…, on the one hand, and 3,4,7,8…, on the other), still higher values were obtained, respectively $r = .84$, $r = .90$, and $r = .78$.

[4] The resistance of the primary TARA index to motivational manipulation remains to be empirically tested. However, it seems a priori unlikely that TARA

effects are solely due to a dearth of motivation: participants in the current research strove hard to complete each TARA as quickly and as accurately as they could, and with the exception of Experiment 1a, were in the presence of another person who was timing them. In addition, external incentives are known to improve performance on simpler tasks, but worsen performance on more complex ones (Wickelgren, 1977). To the extent that completing a TARA honestly constitutes a simpler task, and completing one dishonestly constitutes a more difficult one—as its rationale implies and empirical findings confirm—one might expect stronger motivation to exacerbate rather than attenuate TARA effects.

Table 1
Within-Subject Analysis for Experiments 1a, 1b, 1c, and 2: Means and Standard Deviations of Response Speeds for Honest and Dishonest TARAs, with Inferential Statistics, Effect Size, and Percentage Positive for the Difference between Them

Index	*M*	*SD*	*t*	*d*	*% Positive*
Experiment 1a					
Honest	1029	217	-	-	-
Dishonest	1435	402	-	-	-
Difference	406	318	13.31*	1.28	93%
Experiment 1b					
Honest	1579	322	-	-	-
Dishonest	2090	442	-	-	-
Difference	512	344	10.82*	1.50	98%
Experiment 1c					
Honest	1572	411	-	-	-
Dishonest	2274	604	-	-	-
Difference	702	640	8.43*	1.11	88%
Experiment 2_{I}					
Honest	1539	300	-	-	-
Dishonest	2150	762	-	-	-
Difference	611	568	8.34*	1.09	97%
Experiment 2_{II}					
Honest	1472	310	-	-	-
Dishonest	1888	719	-	-	-
Difference	415	527	6.10*	0.79	90%
Experiment 2_{I+II}					
Honest	1505	296	-	-	-
Dishonest	2019	721	-	-	-
Difference	513	518	7.68*	1.00	93%

Note. N = 109 for Experiment 1a. N = 53 for Experiment 1b. N = 59 for Experiment 1c. N = 58 for Experiment 2. N = 60 for Experiment 2 in all cases.
Means and standard deviations are expressed in milliseconds. In Experiment 2, the subscript "I" refers to the pair of honest and dishonest blocks administered first, the subscript "II" the pair administered second, and "I+II" to their arithmetic average.
* $p < .0001$.

Table 1

Experiments 1a, 1b, 1c, and 2: Correlations between (A) Average Response Speed on Honest and Dishonest TARAs and Within-Subject Difference between Them and (B) Participant Gender, Age, Attitude towards Marmite^{TM}, and Extremity of Attitude

TARA Index	*Gender*	*Age*	*Attitude towards Target*	*Extremity of Attitude*
Experiment 1a				
Honest	.20*	.29**	-.29**	-.04
Dishonest	.00	.24**	-.02	.08
Difference	-.14	.11	.18†	.12
Experiment 1b				
Honest	.18	.15	-.06	-.18
Dishonest	.01	.17	.01	-.19
Difference	-.16	.08	.07	-.07
Experiment 1c				
Honest	.04	.15	.03	-.19
Dishonest	-.15	.27*	.04	-.21
Difference	-.16	.16	-.01	-.07
Experiment 2_{I}				
Honest	.30*	.23†	-.03	.28*
Dishonest	.33*	.33**	.16	.23†
Difference	.28*	.25*	.22†	.17
Experiment 2_{II}				
Honest	.29*	.21	-10	.36**
Dishonest	.32**	.29*	.12	.16
Difference	.26*	.27*	.23†	.01
Experiment 2_{I+II}				
Honest	.34**	.28*	.11	.26*
Dishonest	.32**	.28*	.06	.24†
Difference	.34**	.38*	.09	.25*

Note. N = 109 for Experiment 1a. N = 53 for Experiment 1b. N = 59 for Experiment 1c. N = 58 for Experiment 2.

In Experiment 2, the subscript “I” refers to the pair of honest and dishonest blocks administered first, the subscript “II” the pair administered second, and “I+II” to their arithmetic average.

†< .10. * p < .05. ** p < .01.

Table 3

Between-Subject Analysis for Experiments 1a, 1b, 1c, and 2: Mean Response Speed for Honest and Dishonest TARAs, with Inferential Statistics and Effect Size for the Different between Them

TARA Order	*Honest*	*Dishonest*	*Difference*	*t*	*d*
Experiment 1a					
First	1010	1542	532	7.83^{***}	1.87
Second	1050	1337	286	5.31^{***}	1.06
Experiment 1b					
First	1552	2215	663	6.14^{***}	1.91
Second	1607	1971	364	3.58^{***}	1.00
Experiment 1c					
First	1552	2370	818	5.66^{***}	1.57
Second	1591	2174	583	4.74^{***}	1.36
Experiment 2_{I}					
First	1537	2303	766	5.45^{***}	1.63
Second	1540	1986	446	2.84^{**}	.99
Experiment 2_{II}					
First	1478	2013	534	3.74^{***}	1.12
Second	1466	1754	287	2.02^{*}	0.68
Experiment 2_{I+II}					
First	1507	2158	650	4.74^{***}	1.41
Second	1503	1870	367	2.51^{*}	0.86

Note. In Experiment 1a, alternate *n*s = 57 vs. 52. In Experiment 1b, alternate *n*s = 27 vs. 26. In Experiment 1c, alternate *n*s = 29 vs. 30. In Experiment 2, alternate *n*s = 29 vs. 31. Means are expressed in milliseconds. Values of *t* and *d* are in all cases adjusted to take account of the heterogeneity of variance between honest and dishonest blocks, which is sometimes significant (e.g., in Experiment 1a) and sometimes not (e.g., in Experiment 1c). In Experiment 2, the subscript "I" refers to the pair of honest and dishonest blocks administered first, the subscript "II" the pair administered second, and "I+II" to their arithmetic average.

* $p < .05$. ** $p < .01$. *** $p < .001$.

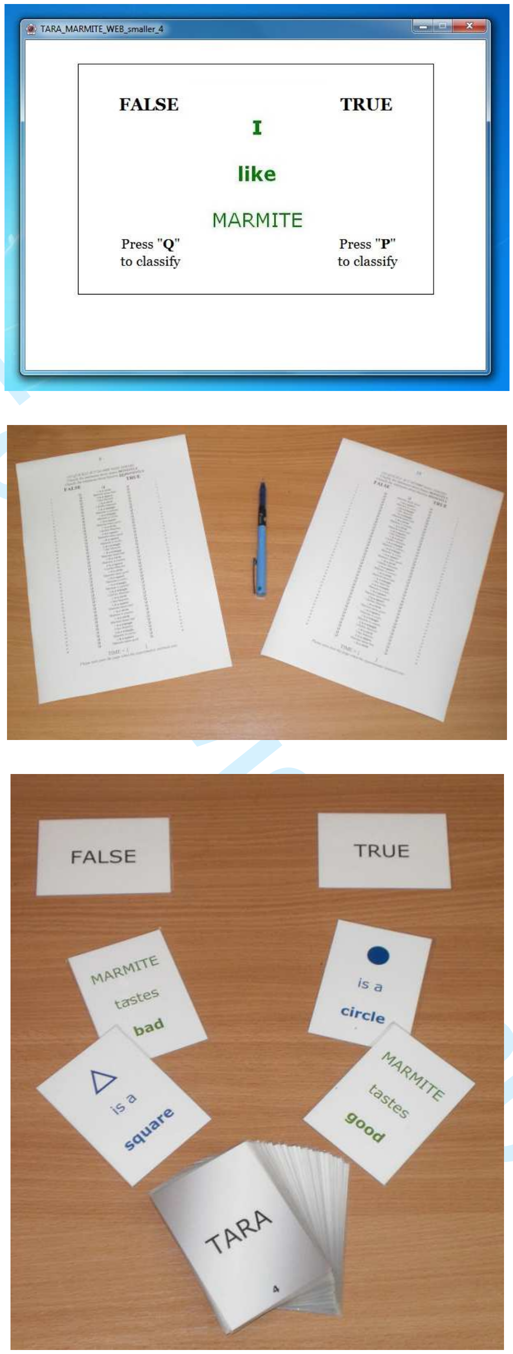
Table 4

Discrimination Analysis for Experiments 1a, 1b, 1c, and 2: Percentages for True Positive, False Positives, True Negatives, and False Negatives, and Overall Accuracy

TARA Order	*True Positive*	*False Positive*	*True Negatives*	*False Negatives*	*Overall Accuracy*
Experiment 1a					
First	73	14	86	27	80
Second	63	27	73	37	68
Experiment 1b					
First	81	11	89	19	85
Second	67	26	74	33	71
Experiment 1c					
First	63	21	79	37	71
Second	62	16	84	38	73
Experiment 2_{I}					
First	71	11	89	29	80
Second	45	19	81	55	63
Experiment 2_{II}					
First	55	21	79	45	67
Second	48	32	68	52	58
Experiment 2_{I+II}					
First	65	17	83	35	74
Second	52	26	74	48	63

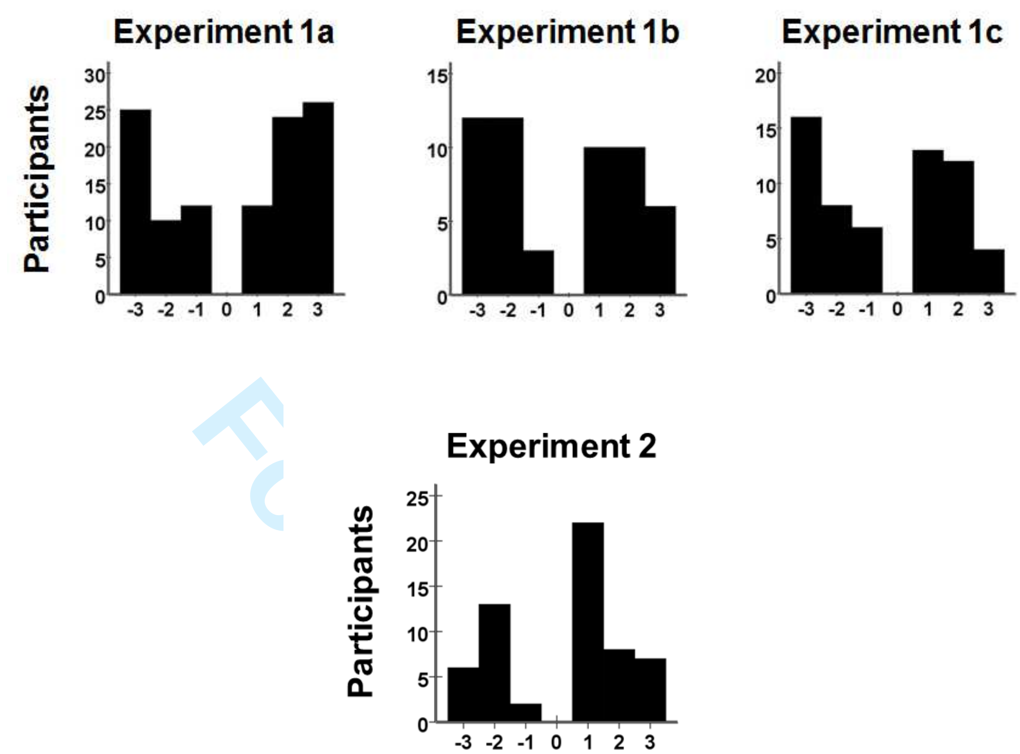
Note. In Experiment 1a, alternate *n*s = 57 vs. 52. In Experiment 1b, alternate *n*s = 27 vs. 26. In Experiment 1c, alternate *n*s = 29 vs. 30. In Experiment 2, alternate *n*s = 29 vs. 31. In Experiment 2, the subscript “I” refers to the pair of honest and dishonest blocks administered first, the subscript “II” the pair administered second, and “I+II” to their arithmetic average.

Figure 1



Images of the materials used in to run TARAs designed to detect deception in attitudes towards *Marmite*™. The top image depicts the computer-based version (Experiment 1a), the middle image the paper-and-pencil version (Experiment 1b), and the bottom image the card-sorting version (Experiment 1c).

Figure 2



Distributions of attitudes towards *Marmite*™ (1 = *I hate Marmite*, 7 = *I love Marmite*) in Experiments 1a, 1b, 1c (top of figure), and Experiment 2 (bottom of figure).

Appendix

Stimuli: Categories and Items

Control Statements (All Experiments)

True	**False**
(1) ○ is a circle	(4) ○ is a triangle
(2) Δ is a triangle	(5) Δ is a square
(3) □ is a square	(6) □ is a circle

Target Statements (Experiments 1a, 1b, and 1c)

True (or False)	**False (or True)**
(7) I like Marmite	(10) I dislike Marmite
(8) Marmite is yummy	(11) Marmite is yucky
(9) Marmite tastes good	(12) Marmite tastes bad

Target Statements (Experiment 2)

True (or False)	**False (or True)**
(7) I like coffee	(10) I dislike coffee
(8) Coffee is yummy	(11) Coffee is yucky
(9) Coffee tastes good	(12) Coffee tastes bad

Stimuli: Pseudo-random order of items

1,12,5,7,6,10,2,8,4,11,3,9,1,10,5,9,3,11,2,7,4,12,6,8,5,10,1,9,3,11,4,8,2,10,6,7,5,12,1,
8,6,12,4,7,2,11,3,9