Detecting concealed language knowledge via response times

Abstract

In the present study, we introduce a response-time-based test that can be used to detect concealed language knowledge, for various potential applications (e.g., espionage, border control, counter-terrorism). In this test, the examinees are asked to respond to repeatedly presented items, including a real word in the language tested (suspected to be known by the examinee) and several pseudowords. A person who understands the tested language recognizes the real word and tends to have slower responses to it as compared to the pseudowords, and, thereby, can be distinguished from those who do not understand the language. This was demonstrated in a series of experiments including diverse participants tested for their native language (German, Hungarian, Polish, Russian; $n = 312$), for second language (English, German; $n = 66$), and several control groups ($n = 192$).

Keywords: deception, language, linguistic profiling, concealed information test, response time

1 Introduction

Methods for discerning the truthfulness of a person who purports to be a native speaker of a language have been recorded throughout history, from at least as early as the 11th century BCE to present day, in various and often crucial scenarios, such as ferreting out infiltrators at battlefronts or verifying asylum claims (Reath 2004; Speiser 1942). Conversely, there is no known test for discerning the truthfulness of a person who *denies* the knowledge of a given language – short of the often repeated anecdote related to the Stroop task, typically recounting how Russian agents during the Cold War were detected based on slower decision on the color of written Russian color words (e.g., Marx and Hillix 1987, p. 410; Peirce and MacAskill 2018, p. 111).

The Stroop task does not actually seem optimal for reliable deception detection,[1] but the espionage scenario does occur in reality and is likely to continue posing a serious threat (Riehle 2020) – undetected cases can be of historical importance (e.g., Black 1987; Kern et al. 2010). Hence, a reliable test for revealing concealed language knowledge could be of great value for intelligence agencies, top-secret research facilities, and other highly confidential organizations. Language tests could also be used for border control, and, in particular, to verify asylum claims: When applicants lack documentation, determining their language knowledge may be used to infer geographical origin (McNamara et al. 2016; Reath 2004).

[1] First, such a simple test is highly susceptible to faking (Boskovic et al. 2018; Hu et al. 2015; Verschuere et al. 2009). Second, the prevalence of color blindness throughout the world is estimated to be around 4-5%, which means that the Stroop effect is diminished in at least 4-5% of the cases to begin with. Third, basic colors are denoted by only a few simple words, thereby, restricting test material – moreover, color words are often similar between languages, and also often known to people who are even just vaguely familiar with a given language. Fourth, it would not be easy to standardize response times of verbal responses, let alone to implement an automatic analysis. Finally, no empirical validation for language detection purposes exists, but the generally moderate effect sizes of the Stroop task (e.g., Homack 2004) foreshadow poor diagnostic efficiency: Very large effect sizes are needed for reliable individual-level lie detection (cf. Suchotzki et al. 2017).

Since the actual origin is often suspected (e.g., Reath 2004), testing for the knowledge of the corresponding language could be used for either preliminary screening or additional support for previous conjectures.[2] Other potential applications include scenarios where a person deceitfully denies the knowledge of a language to (a) avoid cooperation with police or military investigations (i.e., a suspect may deny understanding the language of the authorities, or, alternatively, knowing the local language in a foreign operation); (b) justify, in a legal case, not having understood rules or warnings; (c) claim insurance for language deficiency. Less dramatically, it could also be useful for screening in psycholinguistic experiments in which participants are not supposed to understand a certain language.

Finally, a test for concealed language knowledge could be used to detect not only natural language, but also cants ("cryptolects") and similar coded language: Not only organized crime members, but also terrorists are known to use secret jargon (Koskensalo 2015). Revealing that someone understands such a jargon would clearly warrant serious further scrutiny. Hence, such a test could be valuable for screening national security personnel, passengers at sensitive transport areas, detained terrorist suspects, etc.

All in all, a reliable method for detecting whether or not a person understands a given language could be useful in a variety of high-stakes scenarios. In the present study, we introduce the first method validated for this purpose.

1.1 Task design

In our language detection test, the main items are two real words in the given tested language, and four pseudowords that are graphemically similar to the real words. These items

[2] One method regularly applied in at least a dozen first-world countries is "language analysis for determination of origin" (McNamara et al. 2016). This method is controversial partly due to its questionable validity. For example, the use of a single Pakistani word could lead examiners to believe that the applicant is Pakistani (and, therefore, ineligible for asylum), although this could very well be accidental (Reath 2004, pp. 217-218). Testing for Pakistani language knowledge could provide valuable additional evidence.

are sequentially presented in a random order. The examinee is asked to press a key for each item: One of the two real words is designated (selected randomly in the beginning of the task) as *target* and requires pressing one response key (Key *I* on a standard keyboard), while all other items (the other real word and all four pseudowords) require pressing another response key (Key *E*). The other real word serves as *probe*. We assumed that only those who understand the language would see the probe as saliently different from pseudowords, and that they would respond slower to the probe as compared to pseudowords (which thus serve as *control* items) – and, thereby, based on probe-control (real word versus pseudowords) response time (RT) differences, they can be distinguished from those who do not understand the language.

This expected effect in such a design is supported by a series of related deception detection studies for concealed information (Suchotzki et al. 2017; Verschuere and De Houwer 2011; Verschuere and Meijer 2014), although there is no entirely certain or widely accepted explanation for the underlying mechanism. In our view, it is decisive that the target and probe share at least two interrelated key features (from the perspective of a person who recognizes the probe among the controls): (a) Both target and probe meaning stand out as task-relevant (the target because it requires a different key, and the probe because it pertains to the deception scenario and is thereby semantically salient), and (b) both are, thus, infrequent items compared to the controls – and yet the probe needs to be categorized together with the controls – leading to the response conflict for probes (Lukács and Ansorge 2019; Seymour and Schumacher 2009; Verschuere et al. 2015). It follows that the greater the similarity between target and probe items (relative to the controls), the larger the response conflict (Suchotzki et al. 2018) – hence our choice of real words for both target and probe.

Finally, apart from the main items (probe, target, controls), we included two kinds of fillers that were (same as the general task instructions) always in the language acknowledged

to be understood by the examinee:[3] (a) expressions referring to meaningfulness and genuineness (e.g., "MEANINGFUL," "TRUE," etc.) that had to be categorized with the same key as the target (and, thus, opposite to the probe and the controls), and (b) expressions referring to meaninglessness and fakeness (e.g., "UNTRUE," "FAKE," etc.) that had to be categorized with the same key as the probe and controls. It is assumed (Lukács et al. 2017) that fillers further slow down responses to the probes (when recognized by a person who speaks the language) because the probes have to be categorized together with the semantically incompatible expressions referring to meaninglessness (Nosek et al. 2007; Rosch et al. 1976). In addition, by increasing the complexity of the otherwise excessively simple task, fillers prevent strategically focusing on the target and thereby ignoring, to some extent, the probe and its meaningfulness and relevance (Anderson 1991; Hu et al. 2013; Reber 1989; Verschuere et al. 2015; Visu-Petra et al. 2013).

To establish not only conceptual (task-relevance, frequency) but also semantic correspondence between the probe and the meaningfulness-referring fillers – and thereby further enhance the probe response conflict – the probes (and therefore the targets too) were meaningfulness-referring words as well.

1.2 Study structure

In the first two experiments, the test was performed only by speakers of the tested language (English and German; conducted in behavioral laboratory with university students), demonstrating real word versus pseudoword RT differences. In the subsequent three experiments, nonspeakers were tested too, so that classification accuracy could be assessed (with German, Hungarian, Polish, and Russian, as tested languages; in online experiments sampled from very diverse general populations of the respective countries).

---

[3] A suspect might claim to only speak English, but is suspected to also speak German. In this case, task instructions and fillers in the test are in English. Only the probe and target items are German words (and the controls are German-like pseudowords).

2 Experiments 1 and 2

In Experiments 1 and 2, we tested native German speakers for English knowledge, and for German knowledge, respectively. At the same time, we also examined (a) in Experiment 1, whether meaninglessness-referring words (in the tested language) could serve as better controls than pseudowords, and (b), in Experiment 2, whether pseudowords could serve as better fillers than meaninglessness-referring words (in the instructions' language).

2.1 Method

2.1.1 Participants

Native German speaking students fluent in English participated for course credit at the behavioral laboratory of [university name removed for masked review]. Sample size was decided on by optional stopping using Bayes factor (BF) criterion (BF exceeding 5 for the main within-subject comparison in each given experiment).

In Experiment 1, the initial sample of 60 participants already fulfilled our criteria. The data of 20 participants had to be excluded (6 due to technical issues, 2 due to too low accuracy, 12 for not selecting all four English words correctly during the verification task at the end of the test; as per preregistration), leaving 40 participants (age = 21.2±3.3; 8 male).

In Experiment 2, the initial sample of 55 participants did not fulfill our criteria, hence 20 more participants were invited three times, at which point the criterion was fulfilled with 100 completed tests (as not all invitations were answered). Twenty participants' data had to be excluded (1 due to too low accuracy, 6 for low LexTale score – see below), leaving 93 (age = 21.4±3.1; 38 male).

2.1.2 Procedure

Participants were told about the purpose of the experiment, and they were asked to imagine themselves, during the testing, in a scenario where it would be crucial for them to conceal the knowledge of the tested language (English or German).

The main task, in each test, contained four blocks, each with its own unique set of probe, target, and four controls. Fillers were placed among these items in a random order, but with the restrictions that each of the 9 fillers (3 meaningfulness-referring, 6 meaninglessness-referring) preceded each of the 4 probes, 4 targets, and 16 controls exactly one time. Participants had to press Key *I* when the target appeared, and Key *E* when the probe or a control appeared. Whenever a *meaningfulness-referring* filler appeared, participants had to press the Key *I* (same as for targets), while whenever a *meaninglessness-referring* filler appeared, they had to press Key *E* (same as for probe and controls).

The probes and targets were always real and meaningfulness-referring words in the tested language (English in Experiment 1, with German instructions; and German in Experiment 2, with English instructions). In each block, each probe, target, and control was repeated 18 times. In Experiment 1, for each participant, two blocks had pseudowords as controls, and two other blocks had meaninglessness-referring English words as controls; see Table 1.

Table 1

Item Types Examples for Experiment 1

Item Type	Example 1	Example 2	Correct Key
Target	*meaningful*	*proper*	*#I*
Probe	*genuine*	*true*	*#E*
Control	*onscaft, wrute, sieringlest, deborent*	*unknown, wrong, fake, untrue*	*#E*
Filler-T	*bedeutsam, vertraut, wahr*[4]		*#I*
Filler-NT	*unbedeutend, unvertraut, gefälscht, unbekannt, andere, sonstiges*[5]		*#E*

Note. Each example depicts a possible set of all items in a single block. Example 1 shows possible items in a block with pseudoword controls, while Example 2 shows possible items in a block with meaninglessness-referring controls: The only difference between the two conditions concerned these *controls* – the probe and target items are interchangeable (i.e., they are randomly assigned in each condition from the same pool of words), and the fillers are always identical. Filler-T: "target-side" meaningfulness-referring fillers; Filler-NT: "nontarget-side" meaninglessness-referring fillers.

In Experiment 2, two blocks had, analogously to Experiment 1, meaninglessness-referring English words (instruction language) as fillers to be categorized together with the probe and controls (Key *E*), and two other blocks had pseudoword fillers instead (Table 2).

[4] *Meaningful, familiar, true*.

[5] *Meaningless, unfamiliar, fake, unknown, other, miscellaneous.*

Table 2

Item Types Examples for Experiment 2

Item Type	Example 1	Example 2	Correct Key
Target	*bekannt*[6]	*sinnvoll*[7]	*#I*
Probe	*vertraut*	*bedeutsam*	*#E*
Control	*glätisch, redengig, pauflich, schlinst*	*plaucklos, hokisch, tintzlich, klotselig*	*#E*
Filler-T	*true, meaningful, recognized*		*#I*
Filler-NT	*untrue, fake, foreign, random, unfamiliar, invalid*	*ontreg, dake, saneign, mindaw, unamidiar, imbodal*	*#E*

Note. Example 1 shows possible items in a block with meaninglessness-referring Filler-NT items, while Example 2 shows possible items in a block with pseudoword Filler-NT items: The only difference is in Filler-NT; the probe, target (real German words), and controls (pseudowords) are interchangeable, and Filler-T items are always identical. Filler-T: "target-side" meaningfulness-referring fillers; Filler-NT: "nontarget-side" meaninglessness-referring fillers.

The inter-trial interval randomly varied between 0.5 and 0.8 s. In case of an incorrect response or no response within 1 s, the caption "Inkorrekt!" ("incorrect!") or "Zu langsam!" ("too slow!") appeared in red color, respectively, for 0.5 s, followed by the next trial. The

[6] *Known.*

[7] *Sensible.*

main task was preceded by three short practice rounds that included all items from the upcoming first block, and participants had to repeat any round on which they had too few correct responses in time (for further details see the analogous task in, e.g., Lukács and Ansorge 2019). For analysis, only trials with a correct response between 0.15 s and 1 s were used.

At the end of the language test, as a verification task, participants were shown all probes and controls, and were asked to select the probes (ensuring that they understood the language). The data of those who did not select all four probes correctly were excluded in Experiment 1 (but not in Experiment 2, since all participants were already verified native German speakers). Finally, all participants completed a LexTALE test for English language comprehension (Lemhöfer and Broersma 2012): The data of those with a score below 60% (minimum score for B2 level) were excluded.

To calculate illustrative areas under the curves (AUCs) for probe-control RT mean differences as predictors, we simulated nonspeaker groups for the RT data using 1,000 normally distributed values with a mean of zero and an *SD* derived from the corresponding empirical data as $SD_{real} \times 0.5 + 7$ ms (which has been shown to very closely approximate actual data; Lukács and Specker 2020).

For all five experiments, preregistrations, all testing material (working PsychoPy or JavaScript/HTML codes for each task), the lists of all tested real words and pseudowords in each given language (and detailed description of their origin, creation, and the corresponding selection mechanisms during testing), analysis scripts, collected data, and an online appendix with supplementary analyses are available via https://osf.io/p78u3/?view_only=b581a7a9af7c4a9f91377c920c5731a3. [Temporarily masked preregistration links: Exp. 1: https://osf.io/fq42x/?view_only=6eb975b4221c403187f26ea989e94417, Exp. 2:

https://osf.io/tqr6j/?view_only=93e059a914434c44b180ed79d5602f19, Exp 3: https://osf.io/sg32f/?view_only=6d3f1f66072345eca61372d797b75762, Exp. 4: https://osf.io/2g76c/?view_only=c02550fde63841d6b04bcc6bab12aec9, Exp. 5: https://osf.io/gdk92/?view_only=c64e367ee8e6434ea2080c7d31f6d2e0]

2.2 Results

Large differences (ranging from 43.3 to 156.7 ms) were found between probe and control RTs in both experiments, indicating potential for high classification accuracy; see Table 3. In the within-subject comparison of Experiment 1, blocks with pseudoword controls proved to have 40.11 ms larger probe-control differences than those with meaninglessness-referring controls, 95% CI [19.11, 61.12], $d = 0.61$, 95% CI [0.27, 0.95], $t(39) = 3.86$, $p < .001$, $BF_{10} = 67.54$, indicating higher potential for pseudoword controls. In the within-subject comparison of Experiment 2, blocks with pseudoword fillers and meaninglessness-referring fillers had probe-control differences of only 5.20 ms difference in their magnitudes, 95% CI [–7.33, 17.74], $d = 0.09$, 95% CI [–0.12, 0.29], $t(92) = 0.82$, $p = .412$, $BF_{01} = 6.28$.

Table 3

Response Times and Simulated AUCs in Experiments 1 and 2

	Probe	Control	Target	Filler-NT	Filler-T	P – C	AUC_{sim}
Experiment 1							
Pseudoword	528±70	445±33	569±48	518±51	589±47	83.4±52.0	.928 [.886, .969]
Real word	520±66	477±47	556±46	498±56	589±51	43.3±38.9	.828 [.765, .890]
Experiment 2							
Pseudoword	603±70	446±43	569±45	441±44	592±52	156.7±47.5	.998 [.995, 1]
Real word	606±78	454±42	579±48	498±49	605±54	151.5±57.4	.991 [.982, 1]

Note. Means and *SD*s for individual RT means (ms) for different item types, and for probe-control differences (P – C), and corresponding simulated AUCs (as AUC_{sim}, with 95% CIs in brackets). *Pseudoword* denotes pseudoword controls in Experiment 1, and pseudoword Filler-NT items in Experiment 2; *Real word* denotes meaningfulness-referring controls in Experiment 1, and meaningfulness-referring Filler-NT items in Experiment 2. Filler-T: "target-side" meaningfulness-referring fillers; Filler-NT: "nontarget-side" meaninglessness-referring fillers; AUC: area under the curve.

3 Experiments 3-5

In Experiments 3, 4, and 5, we tested Hungarian natives for German (as a second language) and for Hungarian (native language), and Polish and Russian native speakers for their native languages – and, for all these cases, we also tested respective nonspeaker control groups.

3.1 Method

3.1.1 Participants

Participants for Experiments 3-5 were recruited via the online crowdsourcing platform Prolific (https://www.prolific.co/). The information regarding native and second languages were self-reported by participants on Prolific, and we invited only those who fulfilled our required criteria (e.g., "Hungarian native" and "fluent in German," for testing Hungarian natives for German as a second language).

For Experiments 3 and 4, the preregistered sample sizes were based on the estimated available participants on Prolific. In Experiment 3, the sample was also limited by the actually participating Hungarian participants (hence, collection was stopped, despite not having reached the goal of 50 participants, after 15 days, as preregistered): 41 German speakers and 33 nonspeakers participated out of which 19 had to be excluded (14 for too low German LexTALE score, 5 for too low accuracy), leaving 26 speakers (age = 28.4±6.9; 17 male), and 29 nonspeakers (age = 29.0±8.0; 9 male). Participants were paid 3.10 GBP for the 20-25 min experiment, and a potential 1.55 GBP bonus if they were not detected as understanding German.[8]

In Experiment 4, 50 Hungarian natives and 50 Polish natives participated, both tested for Hungarian as well as Polish (simulating a scenario where two different concealed languages are suspected), hence serving as each other's control groups – out of which 5 had to be excluded (3 for not selecting probes correctly, 2 for too low accuracy), leaving 49 Hungarian (age = 26.2±6.8; 38 male) and 46 Polish speakers (age = 25.1±6.9; 37 male). Participants were paid 4.88 GBP for the 20-25 min experiment, and a potential 0.50 GBP bonus for not having been detected in either language (hence altogether max. 1.00 GBP).

[8] Successful detection for this purpose and for automatic feedback, was based on a $d = 0.3$ (standardized mean difference) between probe and control RTs, a higher level than in previous studies to favor participants (Noordraven and Verschuere 2013).

In Experiment 5, we again used optional stopping (see Footnote 10), which was fulfilled after 130 Russian native speakers (two additions of 30 following an initial 70) and 70 English monolinguals participated, out of which 8 had to be excluded (1 for not selecting probes correctly, 7 for too low accuracy), leaving 124 Russian natives (age = 31.8±10.4 [1 unknown]; 46 male) and 68 English monolinguals (age = 31.1±9.4; 36 male). Participants were paid 3.28 GBP for the 25-30 min experiment, and a potential 0.50 GBP bonus for not having been detected as understanding Russian.

3.1.2 Procedure

The procedure and tasks were the same as in Experiment 1 and 2, unless otherwise noted.

Participants were told about the purpose of the experiment and were asked to imagine themselves, during the testing, in a scenario where it would be crucial for them to conceal any knowledge of the tested language (or both tested languages, in case of Experiment 4).

In Experiment 3, Hungarian native speakers were tested for German and had Hungarian task instructions and fillers. In Experiments 4 and 5 (with participants tested for their native languages), instructions and fillers were in English. Probes and targets were always meaningfulness-referring words in the respective tested language, while the controls were corresponding pseudowords. The target-side fillers were always meaningfulness-referring expressions.

The nontarget-side fillers were meaninglessness-referring expressions in Experiment 3. For a random half of participants in Experiment 4 and for all in Experiment 5, the nontarget-side fillers were meaninglessness-referring expressions in two of the four blocks, but shuffled-letter items in the other two blocks. Preceding filler type change in either case, one short practice round had to be passed before commencing the given block with the new fillers. The six unique probe, target, and controls in each given block served as the basis of the six

nontarget shuffled-letter fillers. The given item's letters were reshuffled for each presentation.[9]

At the end of the test, participants were shown all probes and controls, and were asked to select the probes. As a precaution, in Experiments 4 and 5, the data of those who did not select at least three probes correctly in their native language were excluded. In Experiment 3, participants completed a LexTALE for German, and the data of those with a score below 60% were excluded. (In Experiments 4 and 5, participants completed LexTALE for English for potential exploratory analysis.)

3.2 Results

For all three experiments, AUCs and related data are shown in Table 4.

[9] We hypothesized that shuffled-letter items may be a better mental representation of meaninglessness (nonwords, nonsense words, pseudowords) than words that refer to meaninglessness (and yet are actually meaningful, i.e., existing words), and thereby lead to larger probe-control differences. The BF for comparing the two versions was also used for optional stopping of participant collection. The detailed report on this manipulation, and our related test length analyses, are available via https://osf.io/p78u3/?view_only=b581a7a9af7c4a9f91377c920c5731a3, and are planned to be published in a separate paper. In short, the results seem in line with our preregistered expectations, although their benefit for classification accuracy is limited. They have no important role in the results reported below: The changes in AUCs in particular, depending on different conditions, are relatively small (within 5%).

Table 4

Response Times and AUCs in Experiments 3-5

	Probe	Control	Target	Filler-NT	Filler-T	P – C	AUC	TPR	TNR
Exp. 3 (GE)									
Speaker	519±48	492±39	598±50	573±50	605±53	27.0±25.3	.708 [.571, .845]	.58	.76
Nonspeaker	503±38	493±36	592±47	589±43	609±52	9.4±17.6			
Exp. 4 (HU)									
Speaker	565±70	469±44	577±57	500±48	607±55	95.6±43.8	.992 [.982, 1]	.92	1
Nonspeaker	455±37	461±38	580±39	526±56	587±39	-6.1±13.1			
Exp. 4 (PL)									
Speaker	549±64	489±47	584±53	494±52	601±47	60.0±32.8	.980 [.959, 1]	.91	.98
Nonspeaker	487±54	488±55	595±61	521±60	590±49	-1.3±10.4			
Exp. 5 (RU)									
Speaker	586±79	500±54	616±57	517±54	621±53	85.8±52.1	.939 [.906, .972]	.86	.96
Nonspeaker	497±62	496±61	616±62	539±63	596±57	1.0±17.1			

Note. Means and *SD*s for individual RT means (ms) for different item types, and for probe-control differences (P – C), and, most importantly, corresponding AUCs (95% CIs in brackets). TPR: true positive rates (ratio of correctly detected speakers), TNR: true negative rates (ratio of correctly detected speakers), using arbitrary optimal cutoffs (maximal Youden's index) for classification. Filler-T: "target-side" meaningfulness-referring fillers; Filler-NT: "nontarget-side" meaninglessness-referring fillers. AUC: area under the curve; GE: German; HU: Hungarian; PL: Polish; RU: Russian.

Excluding participants who are suspected of not complying with the requirement is reasonable, but it may be argued that it limits the generalizability of our results and restricts conclusions. Therefore, we exploratorily recalculated AUCs using all participants in all three

experiments without any exclusions. The changes in the AUCs (cf. Table 4) are negligible: .693, 95% CI [.572, .813] (TPR = .59, TNR = .79; 41 speakers, 33 nonspeakers) in Experiment 3; .992, 95% CI [.981, 1] (TPR = .92, TNR = 1; 50 speakers, 50 nonspeakers) in Experiment 4 for the Hungarian language test; .979, 95% CI [.958, 1] (TPR = .90, TNR = .98; 50 speakers, 50 nonspeakers) in Experiment 4 for the Polish language test; .931, 95% CI [.896, .966] (TPR = .85, TNR = .96; 130 speakers, 70 nonspeakers) in Experiment 5.

4 General discussion

First and foremost, we have demonstrated, based on testing three different languages, that our test can detect concealed *native* language knowledge with very high classification accuracy. We have also found strong evidence that the test provides classification accuracy well above chance level for second languages too – although it remains to be shown to what extent the knowledge of specific words and general fluency in the tested language might affect the outcomes.

The complex design of the test offers a number of opportunities for improvement and fine-tuning: As one of many possibilities, in Experiment 1, we have shown that different control items may affect classification accuracy. The test could also be specifically improved for any given language by finding the optimal set of probe, target, and control items. For example, while in our study the probes and targets were assigned randomly (for general proof of concept), it seems likely that pairing close synonyms (as probe and target) would work even better. Testing for concealed language knowledge, as compared to other kinds of deception, is particular in that it does not require experimental setup, such as a mock-crime, to simulate an appropriate scenario. Ground truth is relatively easy to establish (e.g., via a preliminary interview in the examinee's native language), and it is likely that real suspects are no more difficult to detect than experimental participants (Kleinberg, and Verschuere

2016; Suchotzki et al. 2019). Nonetheless, as with all deception detection tests, field testing would be crucial before real-life application.

We invite independent replications and further related research using freely available easy-to-use software for testing and evaluation (Lukács 2019). As explained in the introduction (Section 1), this novel method has wide-ranging potential for screening or for providing additional evidence in various situations – such as spotting spies, criminals, terrorists, or detecting suspicious language-related inconsistencies in legal cases.

References

Anderson, J. R. 1991. The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429. https://doi.org/10.1037/0033-295X.98.3.409

Black, I. 1987. The origins of Israeli intelligence. *Intelligence and National Security*, *2*(4), 151–156. https://doi.org/10.1080/02684528708431920

Boskovic, I., Biermans, A. J., Merten, T., Jelicic, M., Hope, L., and Merckelbach, H. 2018. The modified Stroop Task is susceptible to feigning: Stroop performance and symptom over-endorsement in feigned test anxiety. *Frontiers in Psychology*, *9*, 1195. https://doi.org/10.3389/fpsyg.2018.01195

Homack, S. 2004. A meta-analysis of the sensitivity and specificity of the Stroop Color and Word Test with children. *Archives of Clinical Neuropsychology*, *19*(6), 725–743. https://doi.org/10.1016/j.acn.2003.09.003

Hu, X., Bergström, Z. M., Bodenhausen, G. V., and Rosenfeld, J. P. 2015. Suppressing unwanted autobiographical memories reduces their automatic influences: Evidence from electrophysiology and an Implicit Autobiographical Memory Test. *Psychological Science*, *26*(7), 1098–1106. https://doi.org/10.1177/0956797615575734

Hu, X., Evans, A., Wu, H., Lee, K., and Fu, G. 2013. An interfering dot-probe task facilitates the detection of mock crime memory in a reaction time (RT)-based concealed information test. *Acta Psychologica*, *142*(2), 278–285. https://doi.org/10.1016/j.actpsy.2012.12.006

Kern, G., Schecter, J. L., and Ransom Clark, J. 2010. The trouble with atomic spies. *Intelligence and National Security*, *25*(5), 705–724. https://doi.org/10.1080/02684527.2010.537125

Kleinberg, B., and Verschuere, B. 2016. The role of motivation to avoid detection in reaction time-based concealed information detection. *Journal of Applied Research in Memory*

and Cognition, *5*(1), 43–51. https://doi.org/10.1016/j.jarmac.2015.11.004

Koskensalo, A. 2015. Secret language use of criminals: Their implications to legislative institutions, police, and public social practices. *Sino-US English Teaching*, *12*(7), 497–509. https://doi.org/10.17265/1539-8072/2015.07.005

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

Lukács, G. 2019. CITapp - A response time-based Concealed Information Test lie detector web application. *Journal of Open Source Software, 4*(34), 1179. https://doi.org/10.21105/joss.01179

Lukács, G., and Ansorge, U. 2019. Information leakage in the response time-based Concealed Information Test. *Applied Cognitive Psychology*, *33*(6), 1178–1196. https://doi.org/10.1002/acp.3565

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

Lukács, G., and Specker, E. 2020. Dispersion matters: Diagnostics and control data computer simulation in Concealed Information Test studies. *PLOS ONE, 15*(10), e0240259. https://doi.org/10.1371/journal.pone.0240259

Marx, M. H., and Hillix, W. A. 1987. *Systems and theories in psychology* (4th ed, Vol. 2. McGraw-Hill.

McNamara, T., Van Den Hazelkamp, C., and Verrips, M. 2016. LADO as a language test: Issues of validity. *Applied Linguistics*, *37*(2), 262–283. https://doi.org/10.1093/applin/amu023

Noordraven, E., and Verschuere, B. 2013. Predicting the sensitivity of the reaction time-

based Concealed Information Test: Detecting deception with the Concealed Information Test. *Applied Cognitive Psychology*, *27*(3), 328–335. https://doi.org/10.1002/acp.2910

Nosek, B. A., Greenwald, A. G., and Banaji, M. R. 2007. The Implicit Association Test at Age 7: A methodological and conceptual review. In *Social psychology and the unconscious: The automaticity of higher mental processes* (pp. 265–292. Psychology Press.

Peirce, J., and MacAskill, M. 2018. *Building experiments in PsychoPy*. Sage.

Reath, A. 2004. Language analysis in the context of the asylum process: Procedures, validity, and consequences. *Language Assessment Quarterly*, *1*(4), 209–233. https://doi.org/10.1207/s15434311laq0104_2

Reber, A. S. 1989. Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General*, *118*(3), 219–235. https://doi.org/10.1037/0096-3445.118.3.219

Riehle, K. P. 2020. Russia's intelligence illegals program: An enduring asset. *Intelligence and National Security*, *35*(3), 385–402. https://doi.org/10.1080/02684527.2020.1719460

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., and Boyes-Braem, P. 1976. Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382–439. https://doi.org/10.1016/0010-0285(76)90013-X

Seymour, T. L., and Schumacher, E. H. 2009. Electromyographic evidence for response conflict in the exclude recognition task. *Cognitive, Affective, and Behavioral Neuroscience*, *9*(1), 71–82. https://doi.org/10.3758/CABN.9.1.71

Speiser, E. A. 1942. The Shibboleth Incident (Judges 12:6. *Bulletin of the American Schools of Oriental Research*, *85*, 10–13. https://doi.org/10.2307/1355052

Suchotzki, K., De Houwer, J., Kleinberg, B., and Verschuere, B. 2018. Using more different

and more familiar targets improves the detection of concealed information. *Acta Psychologica, 185*, 65–71.

Suchotzki, K., Kakavand, A., and Gamer, M. 2019. Validity of the reaction time Concealed Information Test in a prison sample. *Frontiers in Psychiatry, 9*, Article 745. https://doi.org/10.3389/fpsyt.2018.00745

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

Verschuere, B., and De Houwer, J. 2011. Detecting concealed information in less than a second: Response latency-based measures. In B. Verschuere, G. Ben-Shakhar, and E. Meijer (Eds.), *Memory detection: Theory and application of the Concealed Information Test* (pp. 46–62). Cambridge University Press.

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

Verschuere, B., and Meijer, E. H. 2014. What's on your mind?: Recent advances in memory detection using the Concealed Information Test. *European Psychologist*, *19*(3), 162–171. https://doi.org/10.1027/1016-9040/a000194

Verschuere, B., Prati, V., and Houwer, J. D. 2009. Cheating the lie detector: Faking in the autobiographical implicit association test. *Psychological Science*, *20*(4), 410–413. https://doi.org/10.1111/j.1467-9280.2009.02308.x

Visu-Petra, G., Varga, M., Miclea, M., and Visu-Petra, L. 2013. When interference helps: Increasing executive load to facilitate deception detection in the Concealed Information Test. *Frontiers in Psychology*, *4*, Article 146. https://doi.org/10.3389/fpsyg.2013.00146