Gáspár Lukács*, Claudia Kawai, Ulrich Ansorge and Anna Fekete

Detecting concealed language knowledge via response times

https://doi.org/10.1515/applirev-2020-0130
Received October 25, 2020; accepted September 16, 2021; published online September 28, 2021

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: concealed information test; deception; language; linguistic profiling; 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

***Corresponding author: Gáspár Lukács**, Department of Cognition, Emotion, and Methods in Psychology, University of Vienna, Vienna, Austria; and Department of Philosophy, Faculty of Psychology, University of Vienna, Liebiggasse 5, A-1010 Vienna, Austria,
E-mail: gaspar.lukacs@univie.ac.at. https://orcid.org/0000-0001-9401-4830
Claudia Kawai and Anna Fekete, Department of Cognition, Emotion, and Methods in Psychology, University of Vienna, Vienna, Austria, E-mail: claudia.kawai@univie.ac.at (C. Kawai), anna.fekete@univie.ac.at (A. Fekete). https://orcid.org/0000-0001-5149-6921 (C. Kawai)
Ulrich Ansorge, Department of Cognition, Emotion, and Methods in Psychology, University of Vienna, Vienna, Austria; Vienna Cognitive Science Hub, University of Vienna, Vienna, Austria; and Research Platform Mediatized Lifeworlds, University of Vienna, Vienna, Austria,
E-mail: ulrich.ansorge@univie.ac.at. https://orcid.org/0000-0002-2421-9942

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). 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

[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.

[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.

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, and so forth.

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 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 vs. 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 and Ansorge 2021; 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 languages (English and German; conducted in behavioral laboratory with university students). In all experiments, including Experiments 1 and 2, participants had to conceal knowledge of only one language – the "*tested*" language –, while another language – the language of the instructions – participants acknowledged to have command of. All five experiments aimed at demonstrating real word versus pseudoword RT differences in the to-be-concealed language, but in Experiments 3–5, nonspeakers were tested too, so that classification accuracy could be assessed (with

[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).

German, Hungarian, Polish, and Russian, as tested languages; in online experiments sampled from very diverse general populations of the respective countries).

2 Experiments 1 and 2

In Experiments 1 and 2, we tested native German speakers for their to-be-concealed English knowledge, and for their to-be-concealed 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 Methods

For all five experiments (the present Experiments 1 and 2, and 3–5 below), 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 (including error rates per item types and conditions) are available as Supplementary Material and via https://osf.io/p78u3/ (direct links to preregistrations: Exp. 1: https://osf.io/fq42x/, Exp. 2: https://osf.io/tqr6j/, Exp 3: https://osf.io/sg32f/, Exp. 4: https://osf.io/2g76c/, Exp. 5: https://osf.io/gdk92).

2.1.1 Participants

Native German speaking[4] students fluent in English participated for course credit at the behavioral laboratory of (university name removed for masked review). The number of participants was decided on by optional stopping using Bayes factor (*BF*; using the default r-scale of 0.707)[5] criterion (BF exceeding 5 for the main within-subject comparison in each given experiment, see below).

[4] In the university's behavioral laboratory system, each potential participant indicates their level of German language comprehension. Only those having selected "native" were invited to participate.

[5] The *BF* is a ratio between the likelihood of the data fitting under the null hypothesis and the likelihood of fitting under the alternative hypothesis (e.g., Jarosz and Wiley 2014). Here, *BF*s are

In Experiment 1, the initial sample of 60 participants already satisfied our *BF* criterion. 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 criterion, 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 data from 93 participants (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 (presented in the instruction language, i.e., German in Experiment 1, English in Experiment 2) 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 (i.e., to-be-concealed) 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. By comparing the average probe-control RT differences between these two kinds of blocks, we will be able to determine which method (pseudoword controls vs. meaninglessness-referring controls) are more effective.

denoted as BF_{10} for supporting the alternative hypothesis, and as BF_{01} for supporting the null hypothesis.

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*[a]		*#I*
Filler-NT	*unbedeutend, unvertraut, gefälscht, unbekannt, andere, sonstiges*[b]		*#E*

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. [a]*Meaningful, familiar, true*. [b]*Meaningless, unfamiliar, fake, unknown, other, miscellaneous*.

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). We reasoned that classifying nontarget filler words from the instructed language together with pseudowords from the tested language could have diminished their joint representation in the control category of Experiment 1 and, hence, the probe-control RT differences. In this case, the RT difference between probes and pseudoword controls may increase when nontarget fillers are also pseudowords (rather than words as in Experiment 1).

Table 2: Item types examples for Experiment 2.

Item type	Example 1	Example 2	Correct key
Target	*bekannt*[a]	*sinnvoll*[b]	*#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*

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. [a]*Known*. [b]*Sensible*.

The inter-trial interval randomly varied between 500 and 800 ms. In case of an incorrect response or no response within 1 s, the caption “Inkorrekt!” (“Incorrect!”) or “Zu langsam!” (“Too slow!”), respectively, appeared in red color for 500 ms, followed by the next trial. The 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 150 ms and 1 s were used.

At the end of the language test, participants were told that they no longer need to conceal their true knowledge of the tested language, and, as a final verification task, they 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)[6] 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).

All analyses were conducted in R (R Core Team 2020; with extension packages by Kelley 2019; Lukács 2020; Morey and Rouder 2018; Robin et al. 2011).

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], $t(39) = 3.86$, $p < 0.001$, $BF_{10} = 67.54$, clearly indicating higher potential for pseudoword controls. In the within-subject comparison of

[6] The AUC is a diagnostic efficiency measure for binary classification that takes into account the distribution of all predictor values (for “receiver operating characteristics”; e.g., Rice and Harris 2005). The AUC can range from 0 to 1, where 0.5 means chance level classification, and 1 means flawless classification (i.e., all guilty and innocent classifications can be correctly made based on the given predictor variable, at a given cutoff point).

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	0.928 [0.886, 0.969]
Real word	520 ± 66	477 ± 47	556 ± 46	498 ± 56	589 ± 51	43.3 ± 38.9	0.828 [0.765, 0.890]
Experiment 2							
Pseudoword	603 ± 70	446 ± 43	569 ± 45	441 ± 44	592 ± 52	156.7 ± 47.5	0.998 [0.995, 1]
Real word	606 ± 78	454 ± 42	579 ± 48	498 ± 49	605 ± 54	151.5 ± 57.4	0.991 [0.982, 1]

Means and SDs 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.

Experiment 2, there was no significant difference between using pseudoword fillers and meaninglessness-referring fillers: The probe-control differences were only nominally larger in case of pseudoword fillers, with 5.20 ms, 95% CI [−7.33, 17.74], $d = 0.09$, 95% CI [−0.12, 0.29], $t(92) = 0.82$, $p = 0.412$, $BF_{01} = 6.28$.

3 Experiments 3–5

In Experiments 3–5, 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 Methods

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 was 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 native speakers 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 German speakers 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.[7]

In Experiment 4, 50 Hungarian and 50 Polish native speakers 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 40–45 min experiment, and a potential 0.50 GBP bonus for not having been detected in either language (hence, altogether max. 1.00 GBP).

In Experiment 5, we again used optional stopping (see Footnote 7), which was fulfilled after 130 Russian native speakers (two additions of 30 following an initial 70 participants) and 70 English monolingual native speakers participated, out of which 8 had to be excluded (1 for not selecting probes correctly, 7 for too low accuracy), leaving 124 Russian native speakers (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 Experiments 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

[7] 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).

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 the to-be-concealed knowledge of the German language 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 (and only participants with sufficient command of the English language were included in the analyses, see below). Probes and targets were always meaningfulness-referring words in the respective tested language, while the controls were corresponding pseudowords. In the beginning of each language test, each participant was shown a list of potential probe and control words (mixed together; in alphabetic order), and was asked to check a box next to any words (up to four) that seemed "meaningful (or in any way very different from the rest of the words)." This served to exclude, on an individual level, potential probes (meaningful words in the tested language) that may have been by accident salient or known to given genuine nonspeakers.

The target-side fillers were always meaningfulness-referring expressions in the instruction language. The nontarget-side fillers were meaninglessness-referring expressions in Experiment 3. For a random half of participants in Experiment 4 and for all examinees 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 any filler type change, 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 for generating the six nontarget shuffled-letter fillers. The given item's letters were reshuffled for each presentation.[8]

At the end of the test, participants were told that they no longer need to conceal their true knowledge of the tested language, 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.

[8] 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/, and are published in a separate paper (Lukács in press), not being relevant to the present one.

Table 4: Response times and areas under the curves 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	0.708 [0.571, 0.845]	0.58	0.76
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	0.992 [0.982, 1]	0.92	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	0.980 [0.959, 1]	0.91	0.98
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	0.939 [0.906, 0.972]	0.86	0.96
Nonspeaker	497 ± 62	496 ± 61	616 ± 62	539 ± 63	596 ± 57	1.0 ± 17.1			

Means and SDs 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 *nonspeakers*), 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.

3.2 Results

For all three experiments, AUCs and related data are shown in Table 4. In all groups, the speakers of a language can be distinguished from nonspeakers well above chance. The classification accuracy (reflected in AUCs, TPRs, and TNRs) is moderate for detecting second language (Experiment 3, German), but very high for detecting native language (Experiment 4, Hungarian, Polish; Experiment 5, 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: 0.693, 95% CI [0.572, 0.813] (TPR = 0.59, TNR = 0.79; 41 speakers, 33 nonspeakers) in Experiment 3; 0.992, 95% CI [0.981, 1] (TPR = 0.92, TNR = 1; 50 speakers, 50 nonspeakers) in Experiment 4 for the Hungarian language test; 0.979, 95% CI [0.958, 1] (TPR = 0.90, TNR = 0.98; 50 speakers, 50 nonspeakers) in Experiment 4 for the Polish language test; 0.931, 95% CI [0.896, 0.966] (TPR = 0.85, TNR = 0.96; 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 (in Experiments 4 and 5, but see also the large probe-control RT differences and the simulated AUCs in Experiment 2). We have also found strong evidence that the test provides classification accuracy well above chance level for second language (Experiment 3, but see also the much larger probe-control RT differences and simulated AUCs in Experiment 1) – 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 very high classification accuracies (e.g., AUCs of 0.94, 0.98, and 0.99) may be surprising in view of the generally moderate ones in previous RT-based concealed information tests (e.g., an average AUC of 0.82 in the meta-analysis by Meijer et al. 2016, and 0.79 in the meta-analysis by Lukács and Specker 2020). However, the recently introduced task design using filler items, as in the present study, has proven very powerful, and has also achieved, for example, similarly high classification accuracy (an AUC of 0.94) when detecting concealed autobiographical details (Lukács et al. 2017, exp. 1). Furthermore, it has been shown that greater probe-control differences are elicited when probes share more semantic features with the target (Suchotzki et al. 2018) or with target-side fillers (Lukács and Ansorge 2021). In the present study's task design, probes and targets were practically as close in semantic relation as it is generally possible for two different words, namely, they were all synonyms. Relatedly, the probes were not merely recognized as different from the controls, but they also directly possessed semantic content (e.g., "true," "meaningful") that conflicted with the key mapping (since probes have to be categorized with one response key, together with meaningless controls, while the meaningful *and* meaningfulness-referring target and fillers have to be categorized with another response key). These factors may have been major contributors to the large effects we found. However, these as well as other potential semantic or lexical influences (e.g., word familiarity, common vs. rare words) would deserve further investigation, and are in fact easy to examine using the framework of our concealed language test, since all main items in the task (probes, targets, controls) may be freely chosen out of all possible words in the given tested language.

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 (meaninglessness-related words vs. pseudowords) 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; e.g., the pair *true-genuine*, or *comprehensible-understood*) 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).

Still, the concealed language test introduced here is not entirely without limitations. Although probably less so than basic color names in the Stroop task (see Section 1), any word in a language may be similar to those in another language (e.g., cognates, false cognates, or loanwords), especially when these languages are related (e.g., same language family, same writing system, etc.), and, therefore, may be recognized by nonspeakers as well. In the same vein, cross-language interference could arise from other languages known to the participant that are not under investigation. This would have to be carefully considered in case of each real-life application of the test. Nonetheless, allowing examinees to exclude words familiar to them (as in our Experiments 3–5) can always mitigate this issue.

Regarding self-reported language knowledge (Experiments 3–5), it is possible that not all participants were honest (in particular, as reflected in the large ratio of below-B2-level LexTALE scores in Experiment 3, many may have falsely or at least exaggeratedly indicated being "fluent" in German). However, this would only mean that our obtained classification accuracies are underestimations, since nonspeakers or nonfluent speakers may have seen no (or less salient) differences between (some of) the real words and the pseudowords (thus possibly explaining the relatively low classification accuracy in Experiment 3 in particular). In case of exclusively genuine speakers of the given languages, the accuracy of our language detection test could be even higher.

Previous studies have found that motivation does not substantially influence the RT-based concealed information test (Kleinberg and Verschuere 2016) and that it is remarkably resistant to faking (Norman et al. 2020; Suchotzki et al. 2021). However, it is not yet certain whether this applies to scenarios of language detection as well: For example, the extreme motivation of spies or terrorists might help them being more successful at altering results. 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.

Acknowledgments: Polish material was provided by Dorota Goc; Russian material was provided by Dinara Talypova and reviewed by Vsevolod Shuvalov, Vadim Kulikov, and Veronika Vishnevskaia – all of whom are thanked for their kind help.
Research funding: Gáspár Lukács has been funded by the OeAW Post-DocTrack Program (grant number: 85065). Claudia Kawai is a recipient of a DOC Fellowship of the Austrian Academy of Sciences at the Department of Cognition, Emotion, and Methods in Psychology at the University of Vienna (grant number: 25068). Additional research grants were provided by the University of Vienna (Förderungsstipendium, JUWI fellowship to C.K.). These funding sources had no role or involvement related to this study.
Author contributions: Concept, study and test designs, software, and analyses by G.L.; test material (words and pseudowords) by C.K. and G.L.; software testing and

administration by G.L., C.K., and F.A.; applications for funding by C.K., G.L., and U.A.; manuscript drafted by G.L., revised by C.K. and U.A., proofread by F.A.

References

Anderson, John R. 1991. The adaptive nature of human categorization. *Psychological Review* 98(3). 409–429.

Black, Ian. 1987. The origins of Israeli intelligence. *Intelligence and National Security* 2(4). 151–156.

Boskovic, Irena, Anita J. Biermans, Thomas Merten, Marko Jelicic, Lorraine Hope & Harald Merckelbach. 2018. The modified Stroop Task is susceptible to feigning: Stroop performance and symptom over-endorsement in feigned test anxiety. *Frontiers in Psychology* 9. 1195.

Homack, Susan. 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.

Hu, Xiaoqing, Angela Evans, Haiyan Wu, Kang Lee & Genyue Fu. 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.

Hu, Xiaoqing, Zara M. Bergström, Galen V. Bodenhausen & J. Peter Rosenfeld. 2015. Suppressing unwanted autobiographical memories reduces their automatic influences: Evidence from electrophysiology and an Implicit Autobiographical Memory Test. *Psychological Science* 26(7). 1098–1106.

Jarosz, Andrew F. & Jennifer Wiley. 2014. What are the odds? A practical guide to computing and reporting Bayes factors. *The Journal of Problem Solving* 7(1). https://doi.org/10.7771/1932-6246.1167.

Kelley, Ken. 2019. *MBESS: The MBESS R package*. R package version 4.5.1. Available at: https://CRAN.R-project.org/package=MBESS.

Kern, Gary, Jerrold L. Schecter & J. Ransom Clark. 2010. The trouble with atomic spies. *Intelligence and National Security* 25(5). 705–724.

Kleinberg, Bennett & Bruno Verschuere. 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.

Koskensalo, Annikki. 2015. Secret language use of criminals: Their implications to legislative institutions, police, and public social practices. *Sino-US English Teaching* 12(7). 497–509.

Lemhöfer, Kristin & Mirjam Broersma. 2012. Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods* 44(2). 325–343.

Lukács, Gáspár. 2019. CITapp - A response time-based Concealed Information Test lie detector web application. *Journal of Open Source Software* 4(34). 1179.

Lukács, Gáspár. 2020. *neatStats: An R Package for neat and painless statistical reporting*. R package version 1.5.1. Available at: https://CRAN.R-project.org/package=neatStats.

Lukács, Gáspár. Prolonged Response Time Concealed Information Test decreases probe-control differences but increases classification accuracy. *Journal of Applied Research in Memory and Cognition*. https://doi.org/10.1016/j.jarmac.2021.08.00, in press.

Lukács, Gáspár & Ulrich Ansorge. 2019. Information leakage in the response time-based Concealed Information Test. *Applied Cognitive Psychology* 33(6). 1178–1196.

Lukács, Gáspár & Ulrich Ansorge. 2021. The mechanism of filler items in the response time concealed information test. *Psychological Research* 85(7). 2808–2828.

Lukács, Gáspár, Bennett Kleinberg & Bruno Verschuere. 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.

Lukács, Gáspár & Eva Specker. 2020. Dispersion matters: Diagnostics and control data computer simulation in Concealed Information Test studies. *PLOS ONE* 15(10). e0240259.

Marx, Melvin H. & William Allen Hillix. 1987. *Systems and theories in psychology*, 4th edn. 2. New York: McGraw-Hill.

McNamara, Tim, Carolien Van Den Hazelkamp & Maaike Verrips. 2016. LADO as a language test: Issues of validity. *Applied Linguistics* 37(2). 262–283.

Meijer, Ewout H., Bruno Verschuere, Matthias Gamer, Harald Merckelbach & Gatthias Ben-Shakhar. 2016. Deception detection with behavioral, autonomic, and neural measures: Conceptual and methodological considerations that warrant modesty. *Psychophysiology* 53(5). 593–604.

Morey, Richard D. & Jeffrey N. Rouder. 2018. *BayesFactor: Computation of Bayes factors for common designs*. R package version 0.9.12-4.2. Available at: https://CRAN.R-project.org/package=BayesFactor.

Noordraven, Ernst & Bruno Verschuere. 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.

Norman, Danielle G., Daniel A. Gunnell, Aleksandra J. Mrowiec & Derrick G. Watson. 2020. Seen this scene? Scene recognition in the reaction-time Concealed Information Test. *Memory & Cognition* 48(8). 1388–1402.

Nosek, Brian A., Anthony G. Greenwald & Mahzarin R. Banaji. 2007. The Implicit Association Test at Age 7: A methodological and conceptual review. *Social psychology and the unconscious: The automaticity of higher mental processes*, 265–292. Hove, UK: Psychology Press.

Peirce, Jonathan & Michael MacAskill. 2018. *Building experiments in PsychoPy*. Los Angeles: Sage.

R Core Team. 2020. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org/.

Reath, Anne. 2004. Language analysis in the context of the asylum process: Procedures, validity, and consequences. *Language Assessment Quarterly* 1(4). 209–233.

Reber, Arthur S. 1989. Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General* 118(3). 219–235.

Rice, Marnie E. & Grant T. Harris. 2005. Comparing effect sizes in follow-up studies: ROC area, Cohen’s d, and r. *Law and Human Behavior* 29(5). 615–620.

Riehle, Kevin P. 2020. Russia’s intelligence illegals program: An enduring asset. *Intelligence and National Security* 35(3). 385–402.

Robin, Xavier, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez & Markus Müller. 2011. pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 12(1). 77.

Rosch, Eleanor, Carolyn B. Mervis, Wayne D. Gray, David M. Johnson & Penny Boyes-Braem. 1976. Basic objects in natural categories. *Cognitive Psychology* 8(3). 382–439.

Seymour, Travis L. & Eric H. Schumacher. 2009. Electromyographic evidence for response conflict in the exclude recognition task. *Cognitive, Affective, & Behavioral Neuroscience* 9(1). 71–82.

Speiser, Ephraim Avigdor. 1942. The shibboleth incident (Judges 12:6). *Bulletin of the American Schools of Oriental Research* 85. 10–13.

Suchotzki, Kristina, Jan De Houwer, Bennett Kleinberg & Bruno Verschuere. 2018. Using more different and more familiar targets improves the detection of concealed information. *Acta Psychologica* 185. 65–71.

Suchotzki, Kristina, Aileen Kakavand & Matthias Gamer. 2019. Validity of the reaction time Concealed Information Test in a prison sample. *Frontiers in Psychiatry* 9. 745.

Suchotzki, Kristina, Burno Verschuere & Matthias Gamer. 2021. How vulnerable is the Reaction Time Concealed Information Test to faking? *Journal of Applied Research in Memory and Cognition* 10(2). 268–277.

Suchotzki, Kristina, Bruno Verschuere, Bram Van Bockstaele, Gershon Ben-Shakhar & Geert Crombez. 2017. Lying takes time: A meta-analysis on reaction time measures of deception. *Psychological Bulletin* 143(4). 428–453.

Verschuere, Bruno & Jan De Houwer. 2011. Detecting concealed information in less than a second: Response latency-based measures. In Bruno Verschuere, Gershon Ben-Shakhar & Ewout Meijer (eds.), *Memory detection: Theory and application of the Concealed Information Test*, 46–62. New York: Cambridge University Press.

Verschuere, Bruno, Bennett Kleinberg & Kalliopi Theocharidou. 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, Bruno & Ewout H. Meijer. 2014. What's on your mind?: Recent advances in memory detection using the Concealed Information Test. *European Psychologist* 19(3). 162–171.

Verschuere, Bruno, Valentina Prati & Jan De Houwer. 2009. Cheating the lie detector: Faking in the autobiographical implicit association test. *Psychological Science* 20(4). 410–413.

Visu-Petra, George, Mihai Varga, Mircea Miclea & Laura Visu-Petra. 2013. When interference helps: Increasing executive load to facilitate deception detection in the Concealed Information Test. *Frontiers in Psychology* 4. 146.

Supplementary Material: The online version of this article offers supplementary material (https://doi.org/10.1515/applirev-2020-0130).