Use the best, ignore the rest

How heuristics allow to tell lie from truth

Verschuere, B.; Lin, Chu-Chien; Huismann, Sara; Kleinberg, B.; Willemse, Marleen; Chong Jia Mei, Emily; Goor, Thierry van; Löwy, Leonie H.S.; Appiah, Obed Kwame; Meijer, Ewout H.

DOI
10.21203/rs.3.rs-1722572/v1

Publication date
2022

Document Version
Submitted manuscript

Use the best, ignore the rest: How heuristics allow to tell lie from truth

Bruno Verschuere (b.j.verschuere@uva.nl)
University of Amsterdam https://orcid.org/0000-0002-6161-4415
Chu-Chien Lin
Sara Huismann
Bennett Kleinberg
Marleen Willemse
Emily Chong Jia Mei
Thierry van Goor
Leonie H. S. Löwy
Obed Kwame Appiah
Ewout Meijer

Research Article

Keywords: Deception, Honesty, Decision Making, Heuristics, Lie detection, Use the best heuristic

Posted Date: June 3rd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1722572/v1

Abstract

The failure to detect deception can have grave consequences in politics, relationships, and the court room (Mazar, Amir, & Ariely, 2008). Yet, people are poor lie detectors and perform barely better than chance (Bond & DePaulo, 2006). Understandably, people struggle with integrating the many putative cues to deception into an accurate veracity judgement. Heuristics simplify difficult decisions by ignoring most of the information and relying instead only on the most diagnostic cues ('use the best, ignore the rest') (Mousavi & Gigerenzer, 2017). We show that people's deception detection ability can be drastically improved by instructing them to rely solely on the single best available cue. We conducted seven studies in which people evaluated honest and deceptive handwritten statements, video transcripts, videotaped interviews, or live interviews. Participants perform at the chance level when they made an unguided judgment free to use any possible cue. But when instructed to rely only on a single cue (detailedness), they were consistently able to detect deception. The heuristics approach even performed at or above the level of an advanced, more resource-intensive credibility assessment method (Nahari, Vrij, & Fisher, 2014). The simplicity and accuracy of the heuristics approach provides for a novel avenue for deception research and hold promise for practice.

Full Text

Deception detection is notoriously difficult

Being able to make a correct lie-truth judgment touches personal life (e.g., infidelity in relationships), legal practice (e.g., detecting false allegations and deceptive denial), and society at large (e.g., does a regime really possess weapons of mass destruction?). However, deception detection is a notoriously difficult task, with people performing barely better than the chance level. A meta-analytic estimate of 24,483 people found their average accuracy in lie-truth discrimination to be only 4% higher than what would be achieved by random guessing. This poor deception detection ability is not restricted to ordinary people, but also found in professionals who routinely engage in deception detection (2). But why is deception detection so challenging?

The needle in the haystack

There exist two prominent explanations as to why people fail at deception detection. First, people rely on the wrong cues. Global surveys have shown that people's beliefs about cues to deception are strong, but wrong (5). A particularly persistent stereotype is that liars avert their eye gaze, despite meta-analytic evidence showing that they do not (7). Second, and arguably more importantly, most cues are – at best – only weakly predictive of deception (6). A meta-analysis found the median standardized effect size (Cohen's *d*) of 88 behavioral cues to be only d=0.10 (7). Put differently, liars and truth tellers display 96% overlap on these behavioral variables. The diagnostic value of most cues was close to zero, and only very few cues – such as richness in detail – show actual promise as cues to deception (8).

The 'many cues' approach

The current approach to improve deception detection is to combine many cues. The Aberdeen Report Judgement Scales, for instance, requires 3 weeks of training for people to be able to use 52 cues for deception detection (9). Rolled out after 9/11, the controversial $900 million program called Screening Passengers by Observation Technique (SPOT) trained airport security personal to screen passengers on 92 cues (10). These training programs, however, have limited success in improving the ability to detect deception (11). We see two main challenges with such a 'many cues' approach. First, as there is only a small number of valid cues, the many cues approach necessarily involves the inclusion of weak cues. Second, people will struggle with combining the many, and often conflicting, cues into a binary veracity judgement (12). Combining many cues will not confuse statistical approaches, but can lead to overfitting, with a considerable drop in accuracy when moving to out-of-sample testing (13). As a radical alternative to the 'many cues' approach, we reasoned the truth may be found in simplicity, and we propose to drop rather than add cues when trying to detect deception (14).

Simple solutions for complex problems

The need to integrate complex information into a binary judgement is not unique to deception detection. Medical doctors, criminal court judges, HR consultants and stockbrokers all face a similar challenge: surgery or medication, guilty or innocent, hire or reject, buy or sell. One counterintuitive way of dealing with an information overload is to simply ignore most of the available information (14, 15). For example, using just two criteria – age and criminal record – allowed to predict the risk of criminal recidivism with the same accuracy as an algorithm that combined 137 criteria (16). And a large-scale study on predicting life outcomes showed that complex computational models did not fare better than domain expert judgments based on just four variables (17). Sometimes, less is more. Would the 'less is more' principle also apply to deception detection when lay people and experts seek to tell lie from truth?

The current Study: Introducing heuristics for deception detection

The use-the-best (and ignore-the-rest) heuristic guides people to rely only on the best available cue. Here, we examine whether this simple heuristic may allow ordinary people to tell lie from truth. In a series of seven studies, we asked people to evaluate honest and deceptive statements. In our control condition, people judged statements directly on veracity (on a continuous scale from completely deceptive to completely honest). While they were free to use any cue they like, and despite financial incentives for accurate performance, we hypothesized that their performance would be close to the chance level (2). In the heuristic condition, we guided people to use only a single cue. Specifically, we used detailedness and the verifiability thereof as heuristics cues as these are the best investigated and likely the most valid cues to deception (4, 7, 8, 18). In the heuristic condition, we therefore guided people to rely on detailedness (Study 2 to 7) and verifiable details (Study 1 to 3), see Table 1.

Study1: Proof of concept

In a pilot (Study1; *n*=39), undergraduate participants evaluated statements of undergraduates who – honestly or dishonestly – described their recent campus activities. Truth tellers had just engaged in

regular campus activities (e.g., call a friend, drink coffee), whereas liars had committed a mock crime: they had stolen an exam but falsely claimed innocent to have been doing campus activities. Participants were randomly allocated in one of two conditions. In the control condition, participants judged these statements on veracity (from -100, totally deceitful to +100, totally truthful). Their judgements did not differ between honest (M = 11.55, SD = 24.51) and deceptive (M = 14.15, SD = 21.86) statements, $t(18)$ = 0.45, p = 0.660, d = 0.10 (95% CI: -0.35; 0.55). In the heuristics condition, we explained people that "Verifiable activities are activities that are recorded (e.g. a security camera), documented (e.g. payment with a debit card or using a smartphone) or an activity with an identifiable witness present" (based on 4), and asked them to use this definition when evaluating the statements on verifiability (also from -100, now meaning totally unverifiable, to +100, now meaning totally verifiable). Their judgements showed large significant differences between honest (M = 33.31, SD = 18.47) and deceptive (M = 14.16, SD = 25.45) statements, $t(19)$ = 4.00, p < .001, d = 0.89, (95% CI: 0.36; 1.41), see Figure 1. Given the considerable and surprising size of this effect, we conducted two preregistered follow-up studies to assess the robustness of the heuristics approach to deception detection.

Study2-3: Registered replication

Study2-3 (combined n =338[1]) followed the same procedure as Study1, but (1) in a larger, crowdsourced sample, (2) with preregistration of the hypotheses and analyses, and (3) an extra heuristics condition that judged the statements on richness in detail. These participants judged the statements on detailedness, using the following definition "the degree to which the message includes details such as descriptions of people, places, actions, objects, events, and the timing of events; the degree to which the message seemed complete, concrete, striking, or rich in details" (7).

The results replicated and firmly supported the pilot findings. The judgements of the control group did not differ between honest (M = 27.12, SD = 28.33) and deceptive (M = 25.43, SD = 26.93) statements, $t(107)$ = 0.58, p = 0.560, d = 0.06 (95% CI: -0.25; 0.13). Note that this poor deception detection ability is seen despite excluding inattentive participants, and despite financial incentives and high self-reported motivation to accurately judge the statements (see Methods). Under the exact same conditions, but now armed with a simple heuristic, people's judgements showed significant and large differences between the deceptive and the honest statements. For detailedness: honest (M = 36.74, SD = 26.05) versus deceptive (M = 15.27, SD = 28.47) statements, $t(102)$ = 11.29, p < 0.001, d = 1.11 (95% CI: 0.86; 1.36), that is honest statements were indeed judged to be higher in detailedness than lies. For verifiable details: honest (M = 38.70, SD = 21.84) versus deceptive (M = 17.00, SD = 25.58) statements, $t(127)$ = 11.89, p < 0.001, d = 1.05 (95% CI: 0.83; 1.26). Honest accounts were judged to be more verifiable than lies. Moreover, the simple heuristics did better than a state-of-the-art, resource intensive deception detection approach by trained coders who coded the statements word for word on (verifiable) details (4).

We translated the heuristics-based judgments into a classification performance by averaging the judgement for each of the 64 statements, for each judgement method. The ROC analysis plots sensitivity against specificity and provides a measure of diagnostic value across all possible cut-off points. As

shown in Table 2, the diagnosticity (expressed as ROC a, ranging from the chance level of 0.50 to 1.00, perfect classification) was above chance for the heuristics condition guiding people to rely on a single cue ($0.71 \leq$ ROC a ≤ 0.75; with the lower bound of the confidence intervals exceeding 0.50), but at chance level for the control condition that allowed considering any possible cues ($0.51 \leq$ ROC a ≤ 0.61; with the 95% confidence intervals including 0.50). Using Youden 's *J* (19), we identified the optimal cut-off point when equally balancing specificity and sensitivity. When we used the data of Study2 to evaluate classification accuracy based on the optimal cut-off derived in Study3 (and vice versa), we found that accuracy was above chance for the heuristic approach (65-70%) and better than the control condition that allowed judges to incorporate any possible cue (50-52%), see Table 2.

Moving from undergraduates (Study1) to crowdsourced participants, Studies 2-3 indicates that our findings are not restricted to a specific sample. But all three studies relied on the same set of statements. To rule out that the effects are artefacts of the stimuli used, we re-tested our findings and found that the benefits of the heuristics conditions over the control condition was seen for each of the 4 subsets of statements that participants had judged (see Supplementary Table 3: https://osf.io/v3kdw/). Still, it is important to assure that our findings generalize beyond the statements used in Study 1 to 3.

Study4. Generalization to other languages and production modes

Studies 1 to 3 established the efficiency of the use-the-best heuristic for deception detection, but all relied on participants evaluating written Dutch statements. To avoid that our findings were specific to a single language or mode of production, we conducted a fourth study (n = 192) with participants evaluating interview transcripts in German. Again, judgments of participants in the control condition did not differ between honest (M = 25.59, SD = 24.83) and deceptive (M = 25.32, SD = 27.85) statements, $t(103) = 0.09$, $p = 0.929$, $d = 0.01$ (95% CI: -0.18; 0.20). But when judging detailedness, there appeared significant and moderate to large differences between honest (M = 19.55, SD = 25.51) and deceptive (M = 1.46, SD = 26.17) statements, $t(87) = 7.06$, $p < .001$, $d = 0.75$ (95% CI: 0.51; 0.99). The heuristic judgements took on average only a minute per statement. This opens the possibility to apply them in real-life situations (e.g., security questioning) where the limited time often prohibits more extensive credibility assessment methods. But before considering real-life applications, we need to assure they are resistant to strong stereotypes about deception.

Study5-6: Towards real-life application.

Study5 (n = 150) addressed a possible confound and paved the way towards the application of heuristics for deception detection. Thus far, participants in the heuristics conditions had not been informed that their judgements served to tell lie from truth. We had been concerned that merely knowing the goal of deception detection may have been enough to activate stereotypes about deceptive behavior (5) thereby overruling the use of the diagnostic cues, and hence diminishing the effectiveness of the heuristics approach. But Study5 mitigates this concern, as the validity of the use-the-best heuristic remained high, even when people knew the goal was deception detection. Crowdsourced participants had been randomly allocated to the non-explicit condition (mimicking the heuristics condition in Studies 1 to 4) or to an

explicit condition. Only participants in the explicit condition were informed that some statements were deceptive and that their goal was to detect deception. We found no evidence that making the goal of deception detection explicit would hamper the validity of the heuristics approach. As before, large significant differences were found between honest (M = 40.56, SD = 23.40) and deceptive (M = 22.69, SD = 25.91) statements when the goal of deception detection was not explicit, $t(73) = 8.74$, $p < .001$, $d = 1.02$ (95% CI: 0.78; ∞). Importantly, this was also the case when the goal of deception detection was explicit: honest (M = 39.56, SD = 24.61) versus deceptive statements (M = 17.33, SD = 28.19), $t(75) = 8.47$, $p < .001$, $d = 0.97$ (95% CI: 0.74; ∞). We conclude that the heuristics approach for deception detection is not easily susceptible to stereotypes about deception. While our heuristic specified what criterion to rely on, it did not instruct the user how to make a decision about the veracity of the statement. In Study6, we added an explicit decision rule, and explored its accuracy under the most challenging conditions – interviewers making decisions on the spot.

In Study6, we explored the applied potential of heuristics for deception detection. Twenty-one deceptive and 23 honest participants were interviewed by four interviewers, who relied on the heuristics approach to make real-time lie-truth decisions. Interviewers judged the statements on detailedness, now on a more user-friendly scale (from 0 = not detailed at all to 10 = very detailed). The participants that were interviewed received a reward if their statement was deemed credible. The simple decision rule was: 'Consider the statement truthful for detailedness scores of 6 and more'. 91% of the truthful statements and 67% of the deceptive statements were correctly classified, with an overall accuracy of 79%. The simple heuristic turned out to be surprisingly accurate.

Study7: Cue diagnosticity matters

Relying on a single cue avoids cognitive overload. But that does not mean any single cue can be validly used in the heuristic approach? To test whether cue diagnosticity matters (12), participants (n = 171) evaluated truthful and deceptive video statements either on a high diagnostic cue (detailedness) or on a low diagnostic cue (the amount of eye gaze aversion), using a scale from 0 to 10. We found strong support for the idea that the cue diagnosticity determines the success of the heuristic approach. As before, we found significant and large differences between honest (M = 7.42, SD = 0.98) and deceptive (M = 5.75, SD = 1.18) statements when participants relied on detailedness, $t(84) = 15.94$, $p < .01$, $d = 1.73$, (95% CI: 1.44; +∞). In contrast, the difference between honest (M = 5.91, SD = 1.16) and deceptive (M = 6.10, SD = 1.17) statements was not significant when participants judged eye gaze aversion, $t(85) = 1.98$, $p = .050$, $d = 0.21$ (95% CI: -0.01; 0.47). Study7 hereby also shows that the success of the heuristics approach cannot be attributed to intuitive decision-making but instead hinges on the diagnosticity of the cue used.

[1] After conducting Study2 we realized it was underpowered for some analyses, and we therefore ran it again with more power (= Study3). As the design is identical, we decided to merge the data of the two studies.

General Discussion

While detecting deception is incredibly important, it is also incredibly difficult. We guided people to only judge the level of detail in the message, and consistently obtained large differences in judgements of lies and truths.

We propose a radical alternative to the trend towards 'many cues' solutions to increase deception detection accuracy (20, 21). Our data show that using many cues is not necessarily needed to increase accuracy. Admittedly, our use-the-best approach is not necessarily restricted to a single cue and could be expanded with other cues. This would, however, require (1) robust evidence for cue validity, and (2) clear guidance on how to combine the cues. Potentially, adding more cues could invalidate the heuristic approach. This risk can be illustrated by a study where people rated statements on 11 cues before making a final lie-truth judgement (22). While users correctly scored truthful statements to be richer in detail than deceptive statements, their final veracity judgements were not above the chance level (22).

To put the accuracy rates found in the current set of studies – 65–79% – into context, we can compare them to other approaches developed to improve deception detection. The cognitive approach advocates active interviewing to increase lie-truth differences by imposing cognitive load, asking unanticipated questions, or encouragements to say more (23). A recent meta-analysis estimated the accuracy of the cognitive approach to be 60%, and, when corrected for publication bias, 55%. Secondly, approaches advocated for and applied in the field – including the Statement Validity Analysis – show accuracy rates in the same range as that observed in our current studies (24). As with any tool applied in real (legal) contexts, the error margin requires caution. But above all, two central findings persisted through a series of seven experiments: first, compared to other approaches, the heuristic approach is a success in its accuracy and efficiency. Second, this paper revives a deception detection approach that has been thought of as a dead-end: human decision-making.

People may not necessarily be poor lie detectors. When judging rich statements about a past event, detailedness provides an easily assessed indicator of truth. The next step is to see whether our findings can be translated to other domains (e.g., the detection of fake news). For now, our findings suggest a simple solution to a complex problem: a rule of thumb may help to find the truth.

References

1. N. Mazar, O. Amir, D. Ariely, The Dishonesty of Honest People: A Theory of Self-Concept Maintenance. J. Mark. Res. **45**, 633–644 (2008).
2. C. F. Bond, B. M. DePaulo, Accuracy of deception judgments. Personal. Soc. Psychol. Rev. (2006) https:/doi.org/10.1207/s15327957pspr1003_2.
3. S. Mousavi, G. Gigerenzer, Heuristics are Tools for Uncertainty. Homo Oeconomicus **34**, 361–379 (2017).

4. G. Nahari, A. Vrij, R. P. Fisher, Exploiting liars' verbal strategies by examining the verifiability of details. Leg. Criminol. Psychol. (2014) https:/doi.org/10.1111/j.2044-8333.2012.02069.x.
5. G. Bogaard, E. H. Meijer, A. Vrij, H. Merckelbach, Strong, but wrong: Lay people's and police officers' beliefs about verbal and nonverbal cues to deception. PLoS One (2016) https:/doi.org/10.1371/journal.pone.0156615.
6. M. Hartwig, C. F. Bond, Why do lie-catchers fail? A lens model meta-analysis of human lie judgments. Psychol. Bull. **137**, 643–659 (2011).
7. B. M. DePaulo, *et al.*, Cues to deception. Psychol. Bull. (2003) https:/doi.org/10.1037/0033-2909.129.1.74.
8. T. J. Luke, Lessons from Pinocchio: Cues to deception may be highly exaggerated. 1–30 (2018).
9. S. L. Sporer, J. Masip, M. Cramer, Guidance to detect deception with the aberdeen report judgment scales: Are verbal content cues useful to detect false accusations? Am. J. Psychol. **127**, 43–61 (2014).
10. S. Weinberger, Airport security: Intent to deceive? Nature **465**, 412–415 (2010).
11. V. Hauch, S. L. Sporer, S. W. Michael, C. A. Meissner, Does Training Improve the Detection of Deception? A Meta-Analysis. Communic. Res. **43**, 283–343 (2016).
12. C. N. H. Street, D. C. Richardson, The focal account: Indirect lie detection need not access unconscious, implicit knowledge. J. Exp. Psychol. Appl. (2015) https:/doi.org/10.1037/xap0000058.
13. B. Kleinberg, Y. van der Toolen, A. Vrij, A. Arntz, B. Verschuere, Automated verbal credibility assessment of intentions: The model statement technique and predictive modeling. Appl. Cogn. Psychol. (2018) https:/doi.org/10.1002/acp.3407.
14. G. Gigerenzer, D. G. Goldstein, Gigerenzer, G., Todd, P.M., & the ABC Research Group. (1999). *Simple Heuristics That Make Us Smart* (1999).
15. A. Tversky, D. Kahneman, Judgment under Uncertainty - Heuristics and Biases Tversky Kahneman 1974. *Science (80-.).* (1974).
16. J. Dressel, H. Farid, The accuracy, fairness, and limits of predicting recidivism. Sci. Adv. **4**, 1–6 (2018).
17. M. J. Salganik, *et al.*, Measuring the predictability of life outcomes with a scientific mass collaboration. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 8398–8403 (2020).
18. M. K. Johnson, C. L. Raye, Reality monitoring. Psychol. Rev. (1981) https:/doi.org/10.1037/0033-295X.88.1.67.
19. W. J. Youden, Index for rating diagnostic tests. Cancer **3**, 32–35 (1950).
20. S. Weinberger, Terrorist "pre-crime" detector field tested in United States. Nature **212** (2011).
21. D. Boffey, EU border "lie detector" system criticised as pseudoscience. Guard. (2018).
22. J. R. Evans, S. W. Michael, Detecting Deception in Non-Native English Speakers. Appl. Cogn. Psychol. (2014) https:/doi.org/10.1002/acp.2990.

23. A. Vrij, R. Fisher, S. Mann, S. Leal, Detecting deception by manipulating cognitive load. Trends Cogn. Sci. **10**, 141–142 (2006).
24. B. Kleinberg, A. Arntz, B. Verschuere, Being accurate about verbal credibility assessment. 1–13 (2019).

Tables

Table 1 Overview of the aim and findings of the 7 studies. Differences in judgement of truthful versus deceptive statements (Cohen´s *d*) when guided to use a single cue (Use the best heuristic). Control condition: Unguided judgements (using any cue; Studies1-4) or a low diagnostic cue (Study7).

Study	Finding	Use the best Heuristic	Control condition
Study1	The Use the best heuristic allows to tell lie from truth	Detailedness: +0.89	Unguided: +0.10
Study2-3	Robustness of Study1 findings in Preregistered Replication	Detailedness: +1.11 Verifiability: +1.05	Unguided: +0.06
Study4	The Use the best heuristic generalizes to novel statements	Detailedness: +0.75	Unguided: +0.01
Study5	Knowing the goal of deception detection did not overrules the Use the best heuristic	Detailedness: +0.97 (goal explicit) Detailedness: +1.02 (goal not explicit)	NA
Study6	Use the best heuristic allows interviewers to make accurate on the spot decisions	Detailedness: +1.86	NA
Study7	The Use the best heuristic critically depends on cue diagnosticity	Detailedness: +1.73	Gaze aversion: +0.21

Table 2. Accuracy in classifying lies from truths for unguided judgments using any possible cue, and for single cue judgements (verifiability, detailedness) for Study2 and 3

	AUC (with 95% CI)		Accuracy	
	Study2	Study3	Cutoff based on Study2 data, applied to Study3 data	Cutoff based on Study3 data, applied to Study2 data
guided ıy cue)	0.61 (0.47; 0.76)	0.53 (0.44; 0.62)	50%	52%
ıgle cue: rifiability	0.72 (0.61; 0.83)	0.75 (0.68; 0.82)	70%	69%
ıgle cue: tailedness	0.72 (0.60; 0.83)	0.71 (0.62; 0.80)	65%	67%

Methods

Study1 (Pilot)

Ethics approval, data, and materials: https://osf.io/z26ar/.

Method

Undergraduates judged handwritten truthful and deceptive alibi statements either on deception (using any cue they like; control condition) or only on verifiability (heuristic condition).

Participants

51 undergraduates of the University of Amsterdam Psychology Department took part in Study1. We excluded seven participants who failed the attention check (see below) and five participants were not Dutch native speakers. Of the 39 remaining participants (31 female, 8 male; *M* age = 19.38 years, *SD* = 1.68), *n* = 19 judged deception (any cue possible) and *n* = 20 judged only the single cue verifiability. Participants received course credits for partaking in the study and the most accurate participant received a 20,- euro bonus.

Procedure

After providing online informed consent, participants were randomly assigned to the deception judgement or the verifiability judgement condition through Qualtrics. In both conditions, the participants were asked to evaluate 16 alibi statements, presented one by one, in a random order. In the verifiability judgement condition there was no mentioning of deception or lie detection.

Judging deception, participants were asked to evaluate “How truthful is this statement?”, on a scale from “totally deceitful” (-100) to “totally truthful” (+100), with the definition of truthfulness provided as “a truthful statement is a statement that is true, honest and adheres to the fact of the situation”. Judging verifiability, participants were asked to evaluate “How verifiable is this statement?”, on a scale from “totally unverifiable” (-100) to “totally verifiable” (+100), with the definition that “verifiable activities are activities that are recorded (e.g. a security camera), documented (e.g. payment with a debit card or using a smartphone) or an activity with an identifiable witness present”. This definition arose from the Verifiability Approach (10), but we simplified it to its essence.

An attention check was embedded among the statements. It looked like another alibi statement but instructed participants to ignore the provided statement and instead answer -47 on the scale. After rating the (real and bogus) statements, a manipulation check asked participants about the basis for the judgments (indicate up to 3 out of the 11 cues from the list they used most as the basis of their judgement; with 3 cues referring to verifiability), a single item asked about motivation to accurately judge the statements (from -100 to +100), and finally participants were asked to provide age, gender, and mother tongue.

Materials

We used 64 alibi statements (32 truthful, 32 deceptive). To avoid item-effects, we created 4 sets of 16 statements (each containing 8 truthful and 8 deceptive statements), and participants were randomly assigned to receive one of the sets. The statements were selected from 72 statements obtained in a previous mock crime study where participants provided a handwritten statement on their whereabouts on campus in the last 15 minutes (1). Participants either truthfully described their activities, or they lied. The lying participants had just enacted the mock theft of an exam but pretended to have been on campus as a regular student. These statements were manually pseudonymized (i.e., all identifiable information including names of persons were changed to plausible alternatives). Content-coding by trained coders (11) showed that the 32 truthful statements (M = 8.28; SD = 8.67) contained more verifiable details than the 32 deceptive statements (M = 3.47; SD = 4.65), d = 0.69 (95% CI: 0.18; 1.19) (see Supplementary Table 3: https://osf.io/v3kdw/). Below is the English translation of one example statement (all original Dutch statements can be found on https://osf.io/z26ar/):

‘*I quietly walked down until the entrance of G/lab. I was in doubt about what to do (stood still for a moment). Then I walked into the corridor of G, saw a cleaner/guy with a cart and read something about using lockers at the UvA. Then I walked to the outside entrance and walked around (back of G) and*

looked at the kind of butterflies that are now there for the light festival. So then walked further around G. Went back inside (second floor lab) and looked for a moment at university pabo, there is a poster next to the door about participating in brain research for money. When I had read that I walked quietly to this research room.'

Main Analyses (Not Preregistered)

The 2 (Judgement Method: Deception vs Verifiability, between-subjects) x 2 (Veracity: Truthful vs Deceptive, within-subjects) mixed ANOVA on the participant's judgment showed the predicted interaction effect between Judgment Method and Veracity, $F(1, 37) = 8.43$, $p = .006$, $\eta^2_p = .19$ (Figure 1). To follow-up on the interaction, we conducted a paired sample t test contrasting judgements for truthful and deceptive statements within each judgement method. Lie-truth differences when judging deception were small and non-significant, $t(18) = 0.45$, $p = 0.660$, $d = 0.10$ (95% CI: -0.35; 0.55)[1], $BF_{01} = 3.85$[2]. In contrast, lie-truth differences when judging verifiability were significant and large, $t(19) = 4.00$, $p < .001$, $d = 0.89$, (95% CI: 0.36; 1.41), $BF_{10} = 45.65$.

Additional analyses

Participants were motivated to provide an accurate judgement (Judging deception: $M = 66.32$; $SD = 29.48$; Judging verifiability: $M = 66.90$; $SD = 32.93$).

The manipulation check showed that participants made judgements in line with the instructions they received in their respective condition: Participants instructed to judge verifiability more often listed cues related to verifiability as the basis of their judgement ($M = 1.75$; $SD = 0.64$) than participants judging deception ($M = 0.68$; $SD = 0.75$), $t(37) = 4.78$, $p < .001$, $d = 1.53$, (95% CI: 0.81; 2.24), $BF_{10} = 669.50$.

Study2-3

Ethics approval, data, and materials: https://osf.io/z26ar/. Moving from the exploratory stages of research to confirmation, we preregistered our hypotheses, analysis plan, and predictions: https://osf.io/z26ar/.

Method

The procedure of Study2-3 followed that of Study1 with 3 main differences. First, we preregistered the hypotheses and statistical analyses. Second, we moved from locally recruited undergraduates to online crowdsourcing (Prolific.co participants with Dutch as first language; Paid 2.50£ for participation, with the 5 best performing participants receiving an additional 5.00£.). Third, we added a heuristic condition that based their judgements on detailedness, using the following definition: 'Degree to which the message includes details such as descriptions of people, places, actions, objects, events, and the

timing of events; the degree to which the message seemed complete, concrete, striking, or rich in details' (2).

There were a few other, minor changes to the Study1 procedure. We provided participants during the initial instructions with a map of the campus. We also changed the manipulation check to an open box, asking participants to describe the cue they relied most on. We added a single item about experienced difficulty of the judgements (from -100 to +100), and – as an additional attention check- after completing all judgments, a surprise multiple-choice question asked about the core of the last statement (e.g., finding a book). Demographics were obtained from Prolific.

Participants

Study2. 142 participants took part in Study2. We excluded 34 participants who failed either of the two attention checks. The 108 remaining participants (39 female, 69 male; *M* age = 29.69 years, *SD* = 10.34; *n* = 30 judging deception, *n* = 39 judging verifiability, and *n* = 39 judging detailedness) mostly had the Dutch nationality (73%; Belgian: 24%; Other: 3%).

Study3. Participants from Study2 could not partake in Study3. 303 participants took part in Study3. We excluded 73 participants who failed either of the two attention checks. Of the 230 remaining participants (107 female, 119 male, 4 missing; *M* age = 30.32 years, *SD* = 10.59; *n* = 77 judging deception, *n* = 89 judging verifiability, and *n* = 64 judging detailedness) 72% had the Dutch nationality (Belgian: 26%; Other: 2%).

Main analyses (preregistered)

Study2. Lie-truth differences when judging deception were small, $d = 0.39$ (95% CI: 0.02; 0.76), and lacked evidential value, $BF_{10} = 1.40$, but were just statistically significant at a significance threshold of 0.05, $t(29) = 2.14$, $p = 0.041$ (see Supplementary Table 3: https://osf.io/v3kdw/). In contrast, lie-truth differences when judging verifiability were again significant and large, $t(38) = 5.06$, $p < .001$, $d = 0.81$, (95% CI: 0.44; 1.17), $BF_{10} = 1854$. This was also true when judging detailedness, $t(38) = 6.61$, $p < .001$, $d = 1.06$, (95% CI: 0.66; 1.45), $BF_{10} = 173709$. However, the 3 (Judgement Method: Deception vs Verifiability vs Detailedness) x 2 (Veracity: Truthful vs Deceptive) mixed ANOVA only showed a main effect of Statement Veracity, $F(1, 105) = 53.03$, $p < .001$, $\eta^2_p = .34$, and not the predicted Judgement Method by Statement Veracity interaction, $F(2, 105) = 1.23$, $p = .297$, $\eta^2_p = .02$. The predicted interaction did not reach significance, but we may have used an underpowered design to uncover the interaction effect. Of note, the lie-truth difference in the control condition happened to be larger than anticipated ($d = 0.39$). We think this is due to sampling error related to the modest sample size (3). We thus ran the study again with more statistical power.

Study3. The 3 (Judgement Method: Deception vs Verifiability vs Detailedness) x 2 (Veracity: Truthful vs Deceptive) mixed ANOVA showed the predicted interaction effect, $F(2, 227) = 31.84$, $p < .001$, $\eta^2_p = .22$. Lie-truth differences when judging deception were small and non-significant, $t(76) = 0.86$, $p = 0.394$, $d =$

0.10 (95% CI: -0.13; 0.32), BF_{01} = 5.60 (see Supplementary Table 3: https://osf.io/v3kdw/). In contrast, lie-truth differences when judging verifiability were again significant and large, $t(88) = 11.60$, $p < .001$, d = 1.23, (95% CI: 0.95; 1.50), $BF_{10} = 2.56 \times 10^{16}$. This was also true when judging detailedness, $t(63) = 9.19$, $p < .001$, d = 1.15, (95% CI: 0.83; 1.46), $BF_{10} = 2.64 \times 10^{10}$.

Additional analyses

Because these analyses are highly similar for Study 2 and 3 (that used the same procedure), we aggregated the data of Study 2 and 3 (n = 338; results for each study separately can be found on https://osf.io/z26ar/).

Preregistered additional analyses.

As a manipulation check of the judgment instructions, we coded the open box responses describing the cue that the participant relied most on. A condition-blind coder scored the responses as referring to verifiability, detailedness or other cues. Agreement with a second condition-blind rater was moderate to high (Study2, all statements double coded: 84% agreement; Study3, one third of statements double coded: 69% agreement). A Chi Square Test on the association between Judgement Method (Judge Deception vs Judge Verifiability vs Judge Detailedness) and Reported Cue Use (Deception vs Verifiability vs Detailedness) indicated that participants reported using the cue they were instructed to use, $\chi^2(4) = 129.08$, $p < .001$, Cramer $V = 0.44$. Participants judging verifiability most often mentioned a cue related to verifiability (89%; Detailedness: 6%; Other: 5%). Participants judging detailedness most often mentioned a cue related to detailedness (44%; Verifiability: 28%; Other: 28%). Participants judging deception most often mentioned other cues (38%; Verifiability: 24%; Detailedness: 37%).

Non-Preregistered additional analyses.

Participants were motivated to provide an accurate judgement (M = 76.99, SD = 23.89; Judging verifiability: M = 81.47; SD = 22.07; Judging deception: M = 75.49; SD = 22.19; Judging detailedness: M = 72.99; SD = 26.90). Participants experienced the task to be moderately difficult (M = 48.29, SD = 38.52; Judging deception: M = 58.75; SD = 36.43; Judging verifiability: M = 43.59; SD = 39.72; Judging detailedness: M = 43.27; SD = 37.27).

Robustness Analyses

Exclusion criteria were preregistered and served to assure that participants paid attention to each statement. But our findings do not hinge on the exclusion criteria. When not excluding any participant, the critical Judgment Method by Statement Veracity interaction was significant for the combined Study 2 and 3 data, $F(2, 442) = 32.77$, $p < .001$, $\eta^2_p = .13$. With large lie-truth differences when judging verifiability: d = 0.95 or detailedness: d = 1.16), but not deception, d = 0.06.

Each participant judged one of 4 series of 16 alibi statements. Splitting the data per stimulus set (bottom rows Supplementary Table 3: https://osf.io/v3kdw/) shows that the benefits of single-cue judgements do not hinge on a specific set of stimuli.

ROC Analyses

For each judgement method, we used the average statement judgement to predict statement veracity. The ROC analysis plots sensitivity against specificity and provides a measure of diagnostic value across all possible cut-off points. The area under the ROC curve varies from 0 to 1 (=perfect classification), with .50 denoting the chance level. As shown in Table 2, classification accuracy was above chance for judgments relying on a single cue (either verifiability or richness in detail), but at chance level for deception judgements.

Using Youden's J (4), we also identify the optimal cut-off point when equally balancing specificity and sensitivity. We used independent validation, the strictest method to avoid data overfitting. Hence, we used the data of Study2 to evaluate classification accuracy based on the optimal cut-off derived in Study3 (and vice versa). Accuracy was poor for veracity judgements, and moderate for the single cue judgements, see Table 2.

Study4

Ethics approval, data, and materials of Study4: https://osf.io/z26ar/. Hypotheses, analysis plan, and predictions were preregistered before the start of data collection: https://osf.io/z26ar/.

Method

Fluent-German crowdsourced-participants judged interview transcripts (see Materials) either on deception or on richness in detail.

Participants

251 participants took part in Study4. We excluded 59 participants who failed both attention checks. The 192 remaining participants (92 females; *M* age = 27.59 years, *SD* = 9.11; *n* = 104 judging deception, *n* = 88 judging detailedness) were Polish (26%), German (13%) or had one of 27 other nationalities. Demographics were obtained from Prolific.

Procedure

After providing informed consent, participants were randomly assigned to the deception judgement or richness in detail judgement condition through Qualtrics. In both conditions, participants were asked to evaluate 13 transcripts (including one bogus transcript used as an attention check, but excluded from

main analyses), presented one by one, in a random order. The instructions for the judgements were the same as for Studies 1 to 3.

The first attention check concerned the bogus transcript, which looked like just another transcript, but with the instruction to ignore the transcript and instead answer -47 on the scale. The second attention check was a surprise recall test after the last transcript, asking to select a unique utterance (e.g., 'forgot the name of the girl I was looking for') in the last transcript among six options. Thereafter, participants indicated their motivation and experienced difficulty, and were asked to list, one-by-one, the cues they had relied on.

Materials

We used 72 transcripts (half truthful, half deceptive). To avoid item-effects, we created 6 sets of 12 transcripts[3], and participants were randomly assigned to receive one of the 6 sets. The statements were selected from (5). Participants in that study were native-German speaking undergraduates who were interviewed about the two tasks they claimed to have been doing in the past half hour. Statements were later transcribed verbatim. We selected the transcripts from participants who had been instructed to consistently lie or tell the truth (i.e., the lie-lie and truth-truth conditions), and used only the 'Find Michelle at the bus stop' task. This task entailed leaving the lab, crossing the campus to the bus stop, trying to find a girl named Michelle (of whom they received a photo), making notes of arriving and leaving buses and returning to the lab within 35 minutes. From the structured interview, we selected only the first response to the interviewer's instruction to describe the task as accurately and in as much detail as possible. Truth tellers described the task they had enacted (trying to find Michelle at the bus stop). Liars also provided a statement about their search to find Michelle, but had not actually enacted that task. We edited the transcripts to correct for spelling errors, but we retained all utterances and filler words (e.g., 'Ehm'). Content-coding of the entire transcripts by trained coders (13) showed that the 36 truthful transcripts (M = 38.06; SD = 15.36) contained more details than the 36 deceptive transcripts (M = 25.72; SD = 9.66), d = 0.96 (95% CI: 0.47; 1.45)[4]. Below is the English translation of one example statement (all original German transcripts can be found on https://osf.io/z26ar/):

'Okay, so after I finished the task with the café I went to the stop at the hospital. I didn't know exactly where it was, so I first meandered through here a bit, asked 'uuh I'm doing a task, can you tell me where the stop is?' And I already thought that it was this one and then I went there. Yes, and then I was supposed to look for Michelle. There were two or three people sitting there, three people sitting there, and then I asked them in Dutch if their name was Michelle. Yes, there was no Michelle there, then I sat there for five minutes, looked to see if maybe some bus was coming by where a Michelle got off, but no bus came by at all. And then I came back here and, yes, I didn't complete the task because I didn't find Michelle'.

Main analyses (preregistered)

The 2 (Judgement Method: Deception vs Richness in detail) x 2 (Veracity: Truthful vs Deceptive) mixed ANOVA showed the predicted interaction effect, $F(1, 190) = 20.09$, $p < .001$, $\eta^2_p = .096$. Lie-truth differences when judging deception were small and non-significant, $t(103) = 0.09$, $p = 0.929$, $d = 0.01$ (95% CI: -0.18; 0.20), $BF_{01} = 9.17$. In contrast, lie-truth differences when judging richness in detail were significant and moderate to large, $t(87) = 7.06$, $p < .001$, $d = 0.75$ (95% CI: 0.51; 0.99), $BF_{10} = 2.53 \times 10^7$.

Additional analyses

Participants were motivated to provide an accurate judgement, *M* = 52.92, *SD* = 39.05 (Judging deception: *M* = 51.78; *SD* = 41.30; Judging richness in detail: *M* = 54.28; *SD* = 36.39) and rated the task moderately difficult, *M* = 42.90, *SD* = 42.94 (Judging deception: *M* = 48.35; *SD* = 44.80; Judging richness in detail: *M* = 36.45; *SD* = 39.95).

Tracking of the time spent per page[5] showed that the average time to read and evaluate a transcript was about a minute, *M* = 57.44 sec., *SD* = 30.81 (Judging deception: *M* = 58.80 seconds; *SD* = 33.82; Judging richness in detail: *M* = 55.82 seconds; *SD* = 26.92).

Study5

Ethics approval, data, and materials of Study5: https://osf.io/z26ar/. Hypotheses, analysis plan, and predictions were preregistered before the start of data collection: https://osf.io/z26ar/.

Method

Participants judged statements on richness in detail, either being explicitly told or not that their judgements served to tell lie from truth. Participants in the explicit condition were told that some statements were deceptive, and that their goal was to detect the deceptive statements. Participants in the non-explicit condition were not given any information about deception or lie detection and merely asked to evaluate the statements.

Participants

166 fluent Dutch-speaking participants (who had not performed in Studies 2 to 4) took part in Study5 on Prolific. We excluded 16 participants who failed both attention checks. The 150 remaining participants (83 females; *M* age = 26.80 years, *SD* = 8.48; *n* = 76 in the explicit condition and *n* = 74 in the non-explicit condition) were Dutch (55.33%), Belgian (31.33%) or had another nationality (13.33%). About half of them (52%) were students. Demographics were obtained from Prolific.

Procedure

After providing informed consent, participants were randomly assigned to the explicit versus non-explicit condition through Qualtrics. In both conditions, participants were asked to evaluate 16 statements (and

an additional bogus statement used as an attention check, but excluded from main analyses), presented one by one, in a random order. Instructions for the non-explicit condition were similar to those used in Studies 1 to 4 (judge detailedness), but in the explicit condition participants were informed (i) that some statements were deceptive and (ii) that their goal was to detect those lies.

The first attention check concerned the bogus statement, which looked like just another transcript, but with the instruction to ignore the transcript and instead answer -47 on the scale. The second attention check was a surprise multiple-choice question after judging the last statement, asking to indicate the core of the last statement (e.g., 'Search for a book') from six options. Thereafter, participants rated motivation and difficulty. Finally, there were two (open box) manipulation checks, asking about the goal of the study and what cues they had relied on.

Materials

The statements were selected from (13) and are the same as those used in Study 1 to 4. We used 64 statements (half truthful, half deceptive). To avoid item-effects, we created 4 sets of 16 statements, and participants were randomly assigned to receive one of the 4 sets.

Main analyses (preregistered)

Using JASP 0.16 and its default settings, the 2 (Goal of lie detection: Explicit vs Non-explicit) x 2 (Veracity: Truthful vs Deceptive) mixed Bayesian ANOVA showed that the data were 2.78 times less likely (BF_{01}) under the model including the interaction than under the model with only the two main effects. Lie-truth differences when judging richness in detail were significant and large when the goal of lie detection was not explicit (as it was in the when participants relied on heuristics in Studies 1 to 4), d = 1.02 (95% CI: 0.78; ∞), BF_{10} = 1.39 x 10^{10} ($M_{truthful}$ = 40.56, SD = 23.40; $M_{deceptive}$ = 22.69, SD = 25.91), but also when the goal of lie detection was made explicit, d = 0.97 (95% CI: 0.74; ∞), BF_{10} = 5.51 x 10^{9} ($M_{truthful}$ = 39.56, SD = 24.61; $M_{deceptive}$ = 17.33, SD = 28.19).

Additional analyses

A condition-blind researcher (MW) coded whether participants mentioned deception or lie detection in the open box answer about the study goal. 56 out of 76 (or 74%) of the participants in the explicit condition mentioned deception or lie detection versus only 2 out of 74 (or 3%) participants in the non-explicit condition, χ^2 (1) = 79.66, p < .001, Cramer´s V = 0.73.

A condition-blind researcher (MW) also coded whether participants mentioned richness in detail in the open box answer about the cues they had relied on. The vast majority of the participants mentioned richness of detail (137 out of 150 or 91.33%).

Participants were motivated to provide an accurate judgement, M = 58.39, SD = 32.81 (Explicit condition: M = 56.07; SD = 35.72; Non-explicit condition: M = 60.78; SD = 29.58) and rated the task as moderately difficult, M = 35.39, SD = 43.42 (Explicit condition: M = 41.07; SD = 42.13; Non-explicit condition: M = 29.57; SD = 44.23).

Robustness check

The exclusion criteria were preregistered and served to assure that participants paid attention to each statement. But our findings do not hinge on the exclusion criteria. When not excluding any participant, the Bayesian ANOVA showed that the data were 2.34 times less likely (BF_{01}) under the model including the interaction than under the model with only the two main effects. And the lie-truth difference was large for the non-explicit condition, d = 1.04 (95%: 0.78; 1.31), as well as the explicit condition, d = 0.97 (95%: 0.71; 1.23).

Study6

Study6 was exploratory and therefore not preregistered. Ethics approval, data, and materials of Study6: https://osf.io/z26ar/.

Method

Undergraduate participants either lied or told the truth in a videotaped interviewed about their whereabouts at the university campus. Immediately after the interview, the interviewers judged the statement on detailedness (using a 0 to 10 scale), with the statement deemed credible for scores of 6 and above. We examined the accuracy of these simple, real-time judgements.

Participants

47 undergraduate participants from the University of Amsterdam took part in return for course credits. Three participants were excluded, two because they did not complete their mission, and one because of suspected intoxication. Of the remaining 44 participants, n = 23 where in the truthful condition (M_{age} = 19.87, SD = 2.70; 43.5% native English speakers; 78% female, 22% male), and n = 21 where in the deceptive condition (M_{age} = 19.48, SD = 1.12; 47.6% native English speakers; 52% female, 43% male, 5% non-binary).

Procedure

We recruited participants who were comfortable to provide a video statement. Through a brief screening via e-mail, we tried to balance our sample and get about half native English and half non-native English speakers. The entire procedure was conducted in English. Upon arrival to the lab, participants were welcomed by a first experimenter[6]. Participants provided written informed consent.

Participants were randomly assigned to the deceptive versus truthful condition and received written instructions for the theft or study location mission, respectively. Participants were asked to paraphrase their mission to the first experimenter to assure it was well-understood. In the deceptive condition, participants first went to a building to find a key, then to a another building to open up a mail box with that key and steal an exam, and finally to a third building to drop the stolen exam. In the truthful condition, participants searched for an appropriate study location in several buildings of the campus, taking flyers with them to proof they visited the designated areas. Participants were asked to return between 25 and 30 minutes.

Upon return to the laboratory, participants were informed that they were suspected of the theft, and were briefly informed about the innocent mission (allowing those in the deceptive condition to create a realistic lie). They were informed that they will be interviewed about their whereabouts in the last half an hour by a second experimenter, and that their statement would be checked on verifiability (see the information protocol of Nahari et al., 2014). A reward in course credits was promised for providing a credible statement, and they were given 10 minutes to prepare the statement.

The participants were then guided to another room, where the second (condition-blind) experimenter would conduct the video interview. After a brief explanation and short small talk (aimed to build rapport), the interviewer asked to describe their whereabouts of the last half an hour in as much detail as possible. To try and get a rich statement, the interviewers encouraged them to fill 10 minutes. The experimenter had been instructed not to interrupt the interviewee during the interviewer, and to only encourage the interviewee to speak (by nodding, 'OK', etc.). When such prompt did not lead the interviewee to say more, a follow-up question was asked (i.e., 'What proves me you are telling the truth?'). Directly after the interview, the same experimenter scored the interview on detailedness from 0 = not detailed at all to 10 = very detailed using the DePaulo et al. (2003) definition ('Degree to which the message includes details such as descriptions of people, places, actions, objects, events, and the timing of events; the degree to which the message seemed complete, concrete, striking, or rich in details').

After the interview, participants were guided back to the first experimenter, asked to take an English language proficiency test, and to honestly answer a few brief questions about the interview experience (single-scale measures of cognitive demand, emotional arousal, motivation, fatigue, and perceived likelihood that statement would be verified; all from -100 not at all to +100)[7]. Participants were thanked and received their credits (with a detailedness score of ≥ 6 by the second experimenter leading to the bonus pay).

Main Analyses (Not Preregistered)

Detailedness judgements for deceptive statements ($M = 7.17$, $SD = 1.34$) were considerably higher than for truthful statements ($M = 4.48$, $SD = 1.57$), $t(42) = 6.16$, $p < .001$, $d = 1.86$ (95%: 1.14; 2.56). The diagnosticity of the detailedness judgements to classify lies from truths was high, ROC = .91 (95% CI: .82; .99). Using the pre-determined cut-off (i.e., 6), 21 out of 23 (91%) truthful statements and 14 out of 21 deceptive statements (67%) were correctly classified. Overall accuracy was 79%. There was a significant

association between statement veracity (truthful versus deceptive statement) and heuristic judgement of veracity (judged truthful versus judged deceptive), χ^2 (1) = 15.94, Cramer *V* = .60.

The six interviewers were trained in coding detailedness using the Verifiability Approach (VA): they got acquainted with the relevant literature, learned the coding scheme https://osf.io/k9e8f/, and practiced the coding in a workshop. Each statement was coded independently by two interviewers who then discussed their coding and came to a consensual scoring. Using this consensus score, the 23 truthful statements (*M* = 15.13; *SD* = 9.67) were found to contain more verifiable details than the 21 deceptive statements (*M* = 6.24; *SD* = 6.09), *t*(42) = 3.61, *d* = 1.09 (95%: 0.45; 1.72).

Study7

Ethics approval, data, and materials of Study7: https://osf.io/z26ar/. Hypotheses, analysis plan, and predictions were preregistered before the start of data collection: https://osf.io/z26ar/.

Method

Participants judged truthful and deceptive videotaped interviews (see Materials) either on a richness in detail or on eye gaze aversion.

Participants

205 participants took part in Study7. We excluded 34 participants who failed both attention checks. The 171 remaining participants (123 female, 44 male, 4 other) had a mean age of 22.07 years (*SD* = 5.03). Eighty-six participants judged detailedness, and 85 judged eye gaze aversion. They had Dutch (31%), English (17.5%) or another (51.5%) language as mother tongue. Their country of origin was The Netherlands (30%), Germany (8%) or one of 40 other nationalities. Participants were rewarded with course credits or 7.50 euro, and the 3 best performing participants received a bonus of 0.50 credit or 5 euro.

Procedure

Participants were recruited via the recruitment portal of the University of Amsterdam. The vast majority of this pool consists of undergraduate students, with the remainder consisting of community members. Participants first provided informed consent, which included the explicit agreement not to download, store or share the video statements. Participants were randomly assigned to judge detailedness or eye gaze behavior through Qualtrics. In both conditions, participants were asked to evaluate 12 videos, presented one by one, in a random order. Instructions for the detailedness judgements were the same as for Studies 1 to 6. For eye gaze aversion, we instructed people to judge 'Looking away', explained that this '… means the person in the video does not maintain eye contact with the interviewer/camera or looking to the side during the interview'.

One attention check asked about the demeanor of the interviewee (i.e., Please answer the following question on the content of the last statement. Which is true? The interviewee was scratching his hair several times, The interviewee was coughing several times, The interviewee was having hiccups several times, The interviewee was holding his nose several times [correct answer], The interviewee was laughing several times), and one attention check asked about the content of the statement (i.e., 'Which of the following persons did in the interviewee mention? Boris Johnson, Joe Biden, Angela Merkel [correct answer], Olaf Scholz, Pope Francis). Thereafter, participants rated motivation and difficulty, and were asked to name the cues they had relied on. Finally, we asked age, native language, gender, country or origin, country of residence, contact detail in order to provide the bonus pay, and whether they opted for money or credits. Finally, participants were debriefed and explained that the interviewees had been instructed to lie vs tell the truth.

Materials

We used 12 video statements obtained in Study6[8]. From the pool of 44 videos, we only used those for which the participants had provided consent to use their video statements in new research, that were below 4 minutes in length after cutting (see below), and where the interviewee was not wearing a face mask. Finally, we selected the videos so that the truthful and deceptive conditions were balanced in native tongue (English vs other). All participants watched the same set of 12 videos (6 truthful, 6 deceptive). From the interview, we cut the initial rapport building phase, and the follow-up question at the end. So we selected the response to the interviewer's instruction to describe in as much detail as possible and trying to fill up the 10 minutes, what the interviewee had done in the last half an hour.

Using the detail count by the trained coders of Study6, the selected 6 truthful videos were found to contain more details (M = 6.17, SD = 1.17) than the selected 6 deceptive videos (M = 4.17, SD = 1.17), d = 1.71 (95% CI: 0.30; 3.03). Using a stopwatch, one team member (OKA) had measured the time that the interviewee looked away from the interviewer/camera. The coding of a random subset (20%) of the statements by second team member (AL) spoke to the reliability of this eye gaze aversion measurement (ICC = .93). That time was converted to the percentage of the entire interview's duration. Eye gaze aversion in the 6 truthful videos (M = 59.83%, SD = 13.70) did not differ from that of the 6 deceptive videos (M = 61.50%, SD = 9.94), d = 0.14 (95% CI: -1.00; 1.27). Thus, the coding confirmed that detailedness, but not eye gaze aversion, is a diagnostic cue to deception.

Main Analyses (preregistered)

Using JASP version 0.16.2, and its default settings, the 2 (Cue: Richness in detail vs Eye Gaze Aversion) x 2 (Veracity: Truthful vs Deceptive) mixed Bayesian ANOVA showed that the data were much more likely (BF_{10} = 5.73 x 10^{23}) under the model that included the interaction as compared to the model that only included the two main effects. As is clear from inspecting Figure 2, cue diagnosticity matters. Lie-truth differences when judging eye gaze aversion were faint, $t(85)$ = 1.98, p = .05, d = 0.21 (95% CI: -0.01; 0.47),

BF_{01} = 1.29. In contrast, lie-truth differences when judging richness in detail were significant and large, $t(84) = 15.94$, $p < .001$, $d = 1.73$, (95% CI: 1.44; +∞), $BF_{10} = 1.79 \times 10^{24}$.

Additional analyses

Participants were motivated to provide an accurate judgement. On a scale from 0 to 10 they rated their motivation $M = 6.14$, $SD = 1.95$ (Judging Eye gaze aversion: $M = 5.97$, $SD = 2.12$; Judging richness in detail: $M = 6.32$, $SD = 1.75$). Also using a 0 to 10 scale, they rated the task to be moderately difficult, $M = 5.06$, $SD = 2.32$ (Judging Eye gaze aversion: $M = 5.22$; $SD = 2.39$; Judging richness in detail: $M = 4.91$; $SD = 2.24$).

Eighty-five percent of the participants of the detailedness condition reported that they relied on detailedness, and 92% of the participants of the eye gaze aversion condition reported that they relied on gaze aversion in their judgment. That participants indeed relied on the instructed cue and can accurately judge cue presence is also apparent from the correlations between the participants judgements and the researcher coding of cue presence. Detailedness judged by the participants correlated strongly with detailedness as assessed by the trained coders, $r = .74$, $p < .01$ (but not with eye gaze aversion as measured with the stopwatch, $r = -.32$, $p = .32$). Eye gaze aversion judged by the participants correlated strongly with eye gaze aversion as measured with the stopwatch, $r = .94$, $p < .001$ (but not with detailedness as assessed by the trained coders, $r = .24$, $p = .45$).

References

1. B. Verschuere, M. Schutte, S. van Opzeeland, I. Kool, The verifiability approach to deception detection: A preregistered direct replication of the information protocol condition of Nahari, Vrij, and Fisher (2014b). *Appl. Cogn. Psychol.* **35**, 308–316 (2021).

2. B. M. DePaulo, *et al.*, Cues to deception. *Psychol. Bull.* (2003) https:/doi.org/10.1037/0033-2909.129.1.74.

3. T. R. Levine, Y. Daiku, J. Masip, The Number of Senders and Total Judgments Matter More Than Sample Size in Deception-Detection Experiments. *Perspect. Psychol. Sci.* (2021) https:/doi.org/10.1177/1745691621990369.

4. W. J. Youden, Index for rating diagnostic tests. *Cancer* **3**, 32–35 (1950).

5. B. L. Verigin, E. H. Meijer, A. Vrij, L. Zauzig, The interaction of truthful and deceptive information. *Psychol. Crime Law* **26**, 367–383 (2020).

[1] Cohen´s *d* is the standardized mean lie-truth difference.

[2] The Bayes Factor BF_{01} expresses how much more likely the data are under the null hypothesis of no lie-truth difference than under the alternative hypothesis of a lie-truth difference. BF_{10} is the inverse of BF_{01}. We report BF_{10} when the data were more likely under the alternative hypothesis than under the null hypothesis.
[3] Due to a programming error 1 of the 6 sets missed 1 (truthful) statement.
[4] Coders counted the number of perceptual, temporal, and spatial details, and we summed these to provide an index of richness in detail. Coding was based on the entire interview.
[5] This Qualtrics feature was only implemented for Study4.
[6] There were 4 experimenters (undergraduates students) who took the role as Experimenter 2, and each interviewed 3 to 14 participants.
[7] Data in full on https://osf.io/z26ar/.
[8] The videos are available from the first author after signing a non-disclosure agreement that stipulates the confidential nature of the videos and that they can only be used for research purposes.

Declarations

Acknowledgements:

General: We thank Abdoullah El Feddali and Vittoria Giannelli for their help with the Pilot Study. We thank Eva Wevers, Romy Louterse, and Jasmine Wong for their help with Study6, and Aaron Lob for his help with Study7.

Funding: Ewout Meijer is supported by the Israel Institute for Advanced Studies.

Author Contributions:

Conceptualization: BV

Methodology: C-C L, SH, MW, LL, TvG, EC, OKA, BV

Investigation: C-C L, SH, MW, LL, TvG, EC, OKA

Statistical analyses: BV, BK

Writing – original draft: BV, EM

Writing – review & editing: BV, EM, BK

Competing interests: Authors declare that they have no competing interests.

Data and Materials availability: All data and materials are publicly available on https://osf.io/z26ar/.

Figures

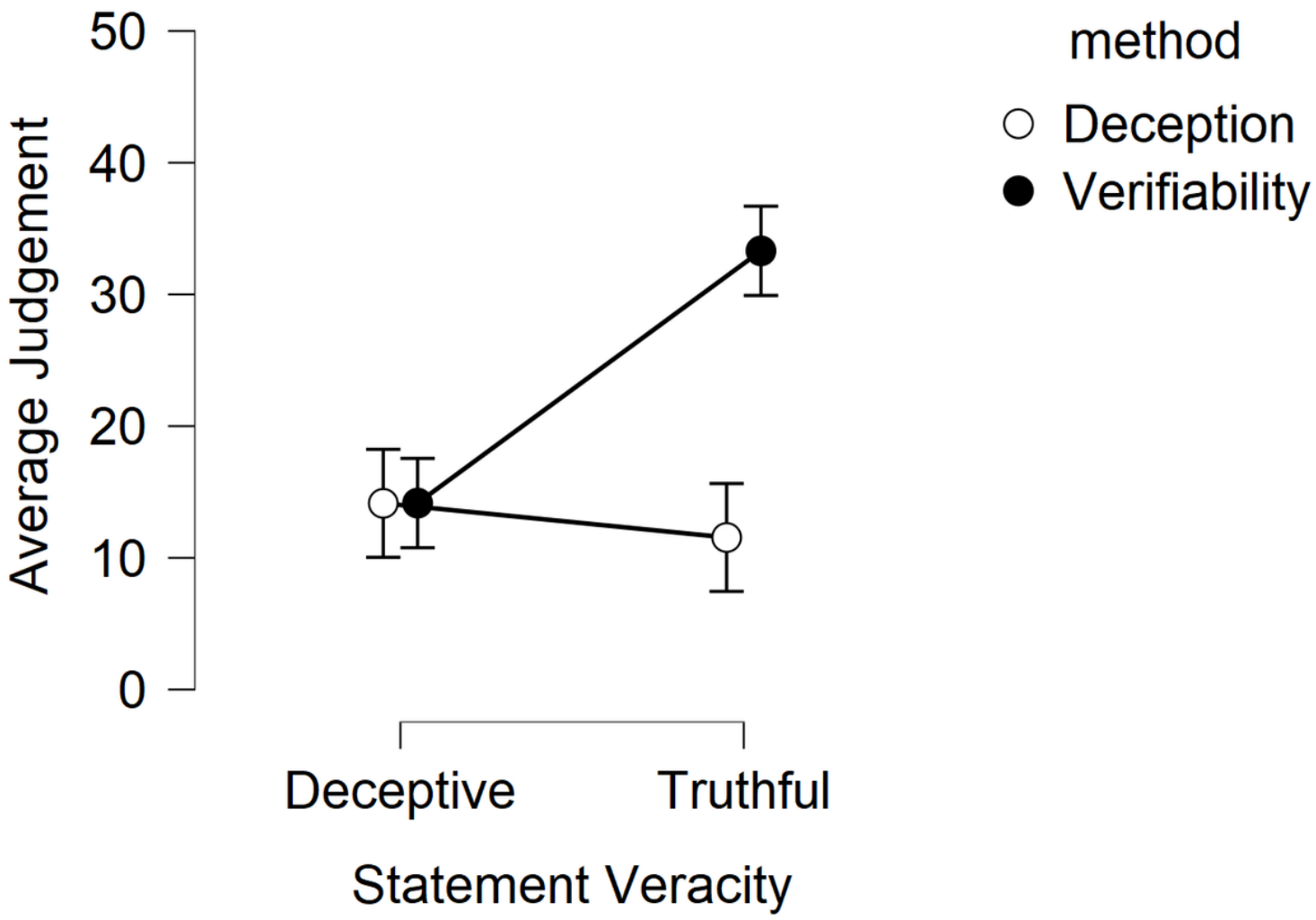

Figure 1

Average judgement of the truthful and deceptive statements when judging deception without guidance (control condition; *n*=19) or verifiability (*n*=20) in Study1. The error bars indicate the standard error of the mean.

Figure 2

Average judgement of the truthful and deceptive statements when judging a high diagnostic (detailedness; *n*=85) or a low diagnostic cue (eye gaze aversion; *n*=86) in Study7. The error bars indicate the standard error of the mean.