

# The use-the-best heuristic facilitates deception detection

---

In the format provided by the  
authors and unedited

## Supplementary Results

### Study1

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 ( $M = 0.68$ ;  $SD = 0.75$ ), two-tailed  $t(37) = 4.78$ ,  $p < .001$ ,  $d = 1.53$ , (95% CI: 0.81; 2.24),  $BF_{10} = 669.50$ .

Key findings of Study1 do not hinge on the applied exclusion criteria. Without any exclusions, the Judgment Method and Veracity was statistically significant,  $F(1, 49) = 8.26$ ,  $p = .006$ ,  $\eta^2_p = .14$ . And the lie-truth difference in the control condition was  $d = 0.08$  whereas it was  $d = 0.76$  for the heuristic condition.

### Study2-3

Supplementary Table1 (top two lines) shows the results for Study 2 and 3 separately.

#### 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:  $M = 43.27$ ;  $SD = 37.27$ ).

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,  $F(2, 442) = 32.77$ ,  $p < .001$ ,  $\eta_p^2 = .13$ . 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 1) shows that the benefits of single-cue judgements do not hinge on a specific set of stimuli.

#### **Study 4**

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:  $M = 36.45$ ;  $SD = 39.95$ ).

In Study4, we also tracked the time spent per page, which 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**

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:  $M = 29.57$ ;  $SD = 44.23$ ).

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,  $\chi^2(1) = 79.66$ ,  $p < .001$ , 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%).

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, two-tailed  $d = 1.04$  (95%: 0.78; 1.31), as well as the explicit condition, two-tailed  $d = 0.97$  (95%: 0.71; 1.23).

## Study6

The continuously scored detailedness judgements by the interviewers 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 these detailedness judgements to classify lies from truths was high,  $ROC = .91$  (95% CI: .82; .99).

There was a significant association between statement veracity (truthful versus deceptive statement) and the binary heuristic judgement of veracity (judged truthful versus judged deceptive),  $\chi^2(1) = 15.94$ , Cramer  $V = .60$ .

The six researchers involved in data collection for Study6 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 ( $M = 6.24$ ;  $SD = 6.09$ ), two-tailed  $t(42) = 3.61$ ,  $d = 1.09$  (95%: 0.45; 1.72).

### **Study7**

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,  $r = .24$ ,  $p = .45$ ).

### **Study8**

Participants were motivated to provide an accurate judgement (Single cue condition:  $M = 6.71$ ;  $SD = 2.15$ ; Multiple cue condition:  $M = 6.98$ ;  $SD = 1.91$ ), and found the task of

moderate difficulty (Single cue condition:  $M = 6.20$ ;  $SD = 2.01$ ; Multiple cue condition:  $M = 6.77$ ;  $SD = 1.86$ ).

In the single cue condition, participants judged the truthful statements ( $M = 6.85$ ,  $SD = 1.22$ ) to be higher on detailedness than the deceptive statements ( $M = 4.85$ ,  $SD = 1.52$ ),  $t(55) = 10.87$ ,  $p < .001$ ,  $d = 1.45$  (95% CI: 1.07, 1.82). In the multiple cue condition, participants judged the truthful statements higher on detailedness ( $M_{\text{truthful}} = 6.83$ ,  $SD = 1.03$  versus  $M_{\text{deceptive}} = 5.41$ ,  $SD = 1.40$ ,  $t(42) = 8.35$ ,  $p < .001$ ,  $d = 1.27$  [95% CI: 0.87, 1.67]), higher on affect ( $M_{\text{truthful}} = 5.16$ ,  $SD = 1.73$  versus  $M_{\text{deceptive}} = 4.41$ ,  $SD = 1.66$ ,  $t(42) = 3.67$ ,  $p < .001$ ,  $d = 0.56$  [95% CI: 0.23, 0.88]), higher on unexpected complications ( $M_{\text{truthful}} = 4.99$ ,  $SD = 1.48$  versus  $M_{\text{deceptive}} = 4.44$ ,  $SD = 1.47$ ,  $t(42) = 2.52$ ,  $p = .016$ ,  $d = 0.38$  [95% CI: 0.07, 0.69]), yet lower on admitting lack of memory ( $M_{\text{truthful}} = 3.81$ ,  $SD = 1.52$  versus  $M_{\text{deceptive}} = 4.55$ ,  $SD = 1.66$ ,  $t(42) = 3.57$ ,  $p < .001$ ,  $d = -0.54$  [95% CI: -0.22, -0.86]) than the deceptive statements.

Because of our preregistration and sample size we reported the student  $t$ -test. Yet, accuracy data were left-skewed. The non-parametric Mann-Whitney test also confirmed that the single cue condition outperformed the many cues condition,  $U = 1494.50$ ,  $p = 0.018$ ,  $r_{pb} = 0.241$ .

Exclusion criteria were preregistered, and resulted in a substantial proportion of the participants being excluded (i.e., 47 out of 146 participants of 32%). Importantly, when not excluding any participant and running the analyses on the full sample of  $n=146$ , the key hypothesis was again confirmed: The one-tailed independent sample  $t$ -test showed again that participants accuracy (%) in telling lies from truths was higher when relying on a single cue (58.95%) than when relying on multiple cues (53.46%),  $t(144) = 2.790$ ,  $p = .003$ ,  $d = 0.408$  (95% CI: 0.19,  $+\infty$ ),  $BF_{10}=11.93$ . In fact, the  $p$  value, effect size, and Bayes Factor showed

that the effect was even stronger without the exclusions. This indicates that the attention check unnecessarily excluded participants (reason why we changed it in Study9).

## Study9

Participants were motivated to provide an accurate judgement (Single cue condition:  $M = 7.41$ ;  $SD = 2.08$ ; Multiple cue condition:  $M = 7.31$ ;  $SD = 2.07$ ), and found the task of moderate difficulty (Single cue condition:  $M = 6.10$ ;  $SD = 2.36$ ; Multiple cue condition:  $M = 6.45$ ;  $SD = 2.03$ ).

In the single cue condition, participants judged the truthful statements ( $M = 6.35$ ,  $SD = 1.17$ ) to be higher on detailedness than the deceptive statements ( $M = 4.41$ ,  $SD = 1.23$ ),  $t(194) = 17.51$ ,  $p < .001$ ,  $d = 1.25$  (95% CI: 1.06, 1.44). In the multiple cue condition, participants judged the truthful statements higher on detailedness ( $M_{\text{truthful}} = 6.45$ ,  $SD = 1.24$  versus  $M_{\text{deceptive}} = 4.65$ ,  $SD = 1.46$ ,  $t(186) = 16.49$ ,  $p < .001$ ,  $d = 1.21$  [95% CI: 1.02, 1.39]), higher on affect ( $M_{\text{truthful}} = 5.45$ ,  $SD = 1.37$  versus  $M_{\text{deceptive}} = 4.63$ ,  $SD = 1.47$ ,  $t(186) = 9.39$ ,  $p < .001$ ,  $d = 0.69$  [95% CI: 0.53, 0.84]), higher on unexpected complications ( $M_{\text{truthful}} = 5.34$ ,  $SD = 1.40$  versus  $M_{\text{deceptive}} = 4.26$ ,  $SD = 1.47$ ,  $t(186) = 8.27$ ,  $p < .001$ ,  $d = 0.60$  [95% CI: 0.45, 0.76]), and higher on spontaneous corrections ( $M_{\text{truthful}} = 5.40$ ,  $SD = 1.43$  versus  $M_{\text{deceptive}} = 4.56$ ,  $SD = 1.46$ ,  $t(186) = 11.29$ ,  $p < .001$ ,  $d = -0.83$  [95% CI: 0.66, 0.99]) than the deceptive statements.

Because of our preregistration and sample size we reported the student  $t$ -test. Yet, accuracy data were left-skewed. The non-parametric Mann-Whitney test also confirmed that the single cue condition outperformed the many cues condition,  $U = 22965.50$ ,  $p < 0.001$ ,  $r_{pb} = 0.260$ .

Exclusion criteria were preregistered. The key effect does not hinge on the applied exclusion criteria: A one-tailed independent sample t-test on the full sample of  $n = 405$  also showed that participants achieved higher accuracy in telling lies from truths when relying on a single cue (65.92%) than when relying on multiple cues (58.46%),  $t(403) = 4.942$ ,  $p < .001$ ,  $d = 0.491$  (95% CI: 0.325,  $+\infty$ ),  $BF_{10} = 22068$ .

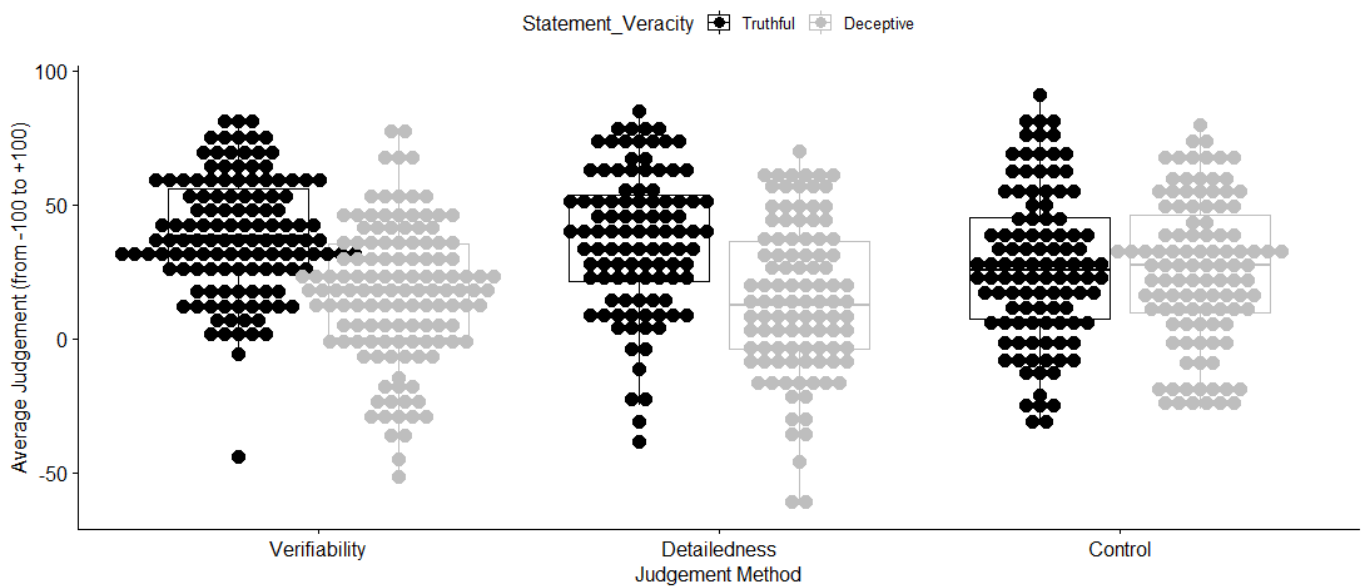
## Supplementary Methods

### Study2-3

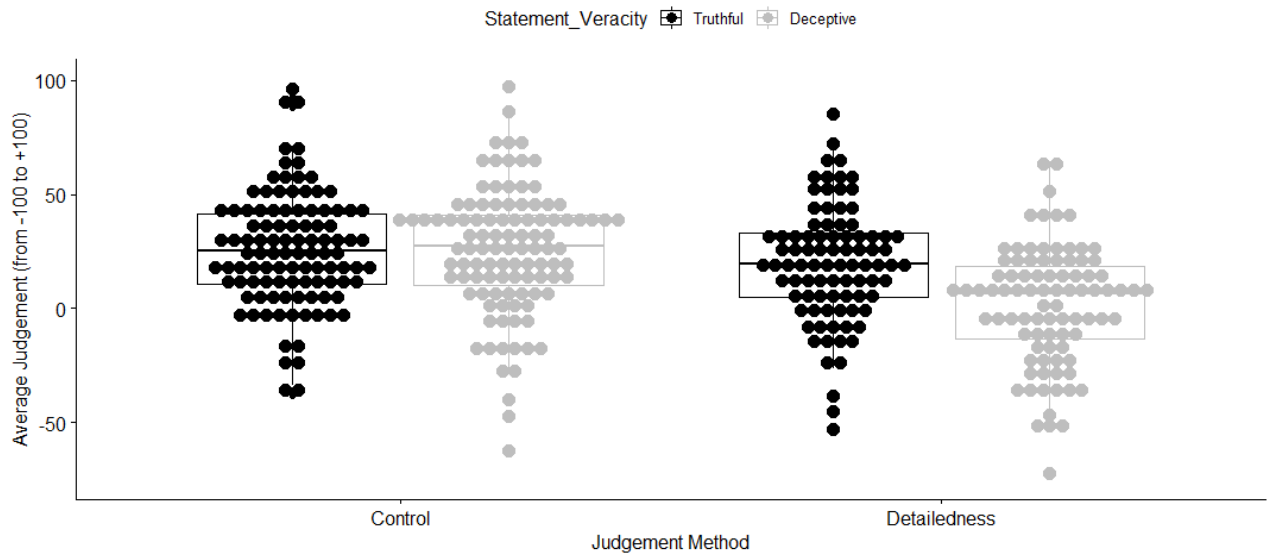
There were a few other, minor changes to the Study1 procedure: We provided participants during the initial instructions with a map of the campus. We changed the manipulation check to an open box, asking participants to describe the cue they relied most on. We added a single item about experienced difficulty 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 including sex (options: male, female) were obtained from Prolific.

## Supplementary Figures

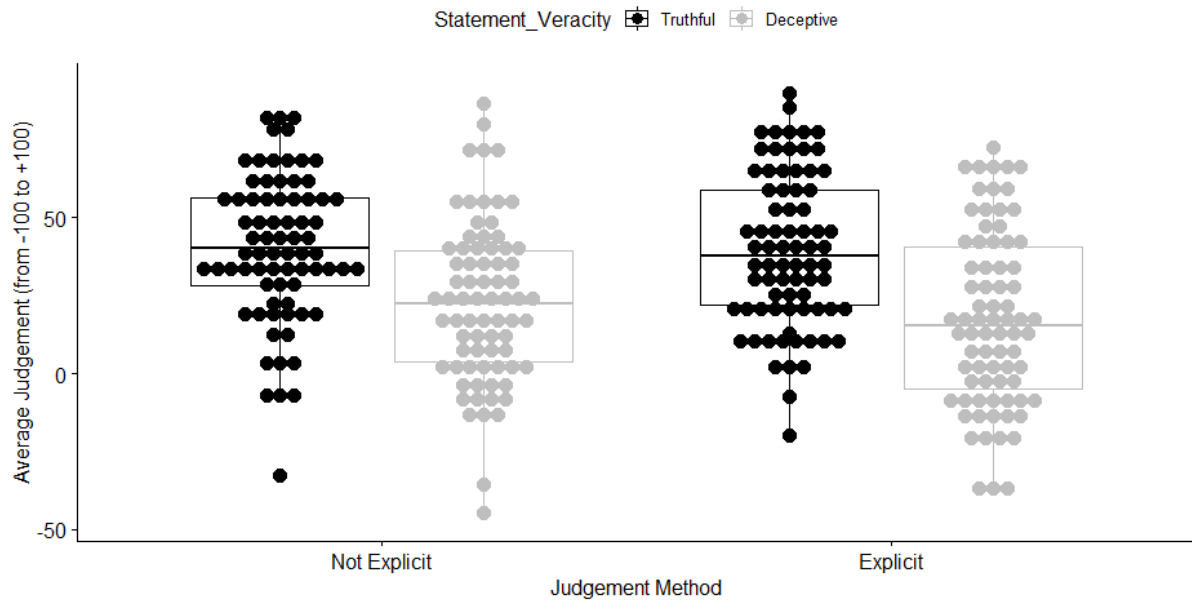
Supplementary Figure 1. Average judgement of the truthful and deceptive statements when judging deception without guidance (control condition;  $n=107$ ), detailedness ( $n=103$ ), or verifiability ( $n=128$ ) in Study2-3. The boxplot shows the median (the midline of the box), the interquartile range (the outer borders of the box), and extreme values (i.e., values larger than 1.5 times the interquartile range).



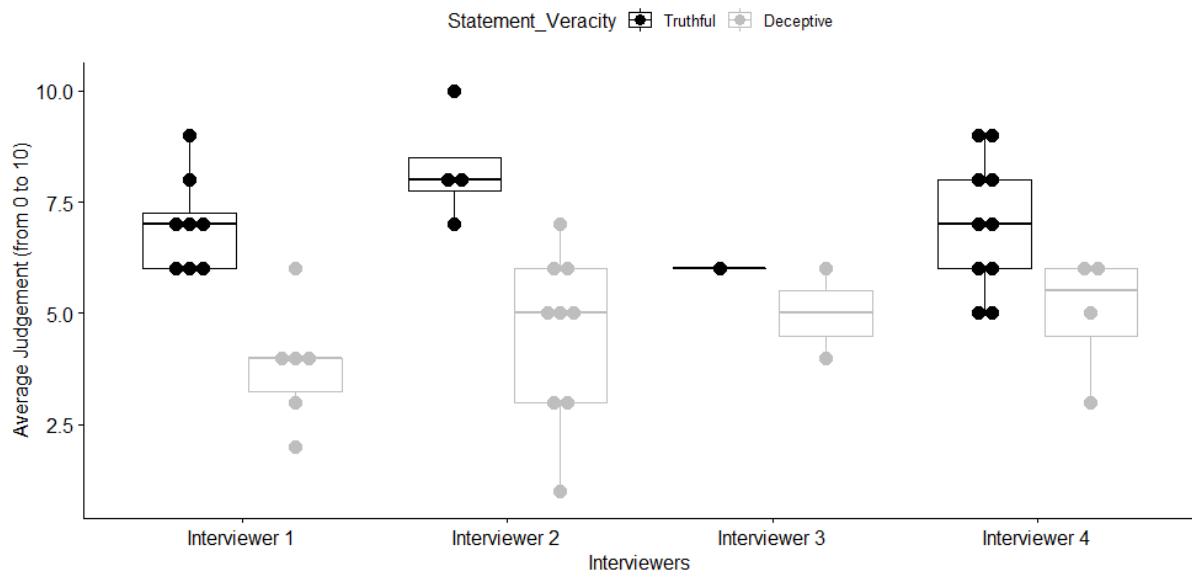
Supplementary Figure 2. Average judgement of the truthful and deceptive statements when judging deception without guidance (control condition;  $n=104$ ) or detailedness ( $n=88$ ) in Study4. The boxplot shows the median (the midline of the box), the interquartile range (the outer borders of the box), and extreme values (i.e., values larger than 1.5 times the interquartile range).



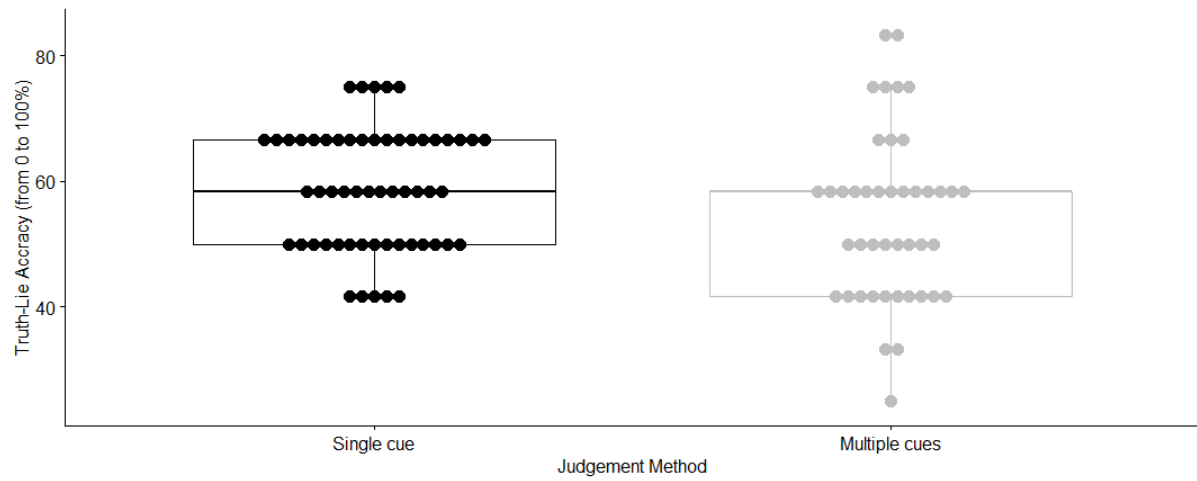
Supplementary Figure 3. Average detailedness judgement of the truthful and deceptive statements when the goal of lie detection was made explicit ( $n=76$ ) or not ( $n=74$ ) in Study5. The boxplot shows the median (the midline of the box), the interquartile range (the outer borders of the box), and extreme values (i.e., values larger than 1.5 times the interquartile range).



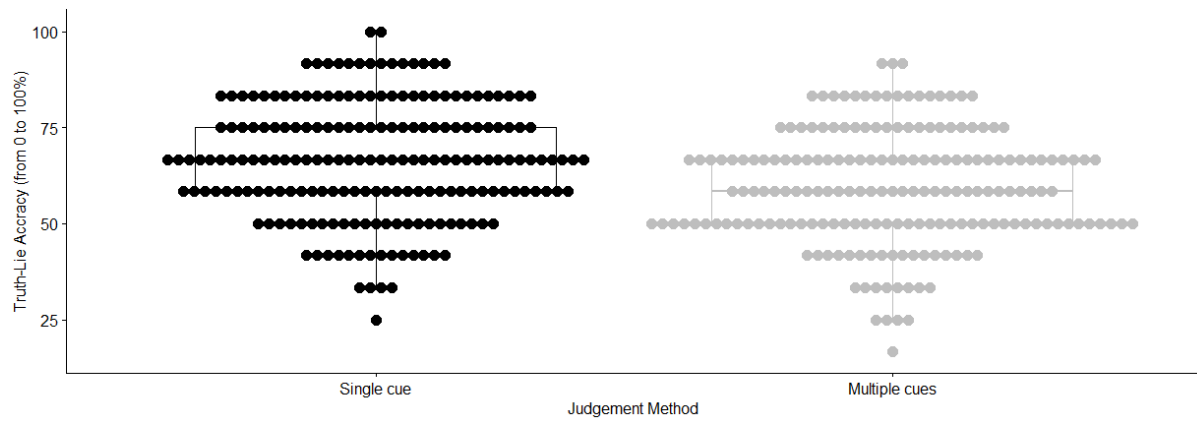
Supplementary Figure 4. Average detailedness judgement of the 23 truthful and 21 deceptive statements by the 4 interviewers in Study6. The boxplot shows the median (the midline of the box), the interquartile range (the outer borders of the box), and extreme values (i.e., values larger than 1.5 times the interquartile range).



Supplementary Figure 5. Average accuracy (%) in lie-truth discrimination for the single cue ( $n=56$ ) and the many cue ( $n=43$ ) conditions in Study8. The boxplot shows the median (the midline of the box), the interquartile range (the outer borders of the box), and extreme values (i.e., values larger than 1.5 times the interquartile range).



Supplementary Figure 6. Average accuracy (%) in lie-truth discrimination for the single cue and ( $n=195$ ) the many cue conditions ( $n=187$ ) in Study9. The boxplot shows the median (the midline of the box), the interquartile range (the outer borders of the box), and extreme values (i.e., values larger than 1.5 times the interquartile range).



## Supplementary Tables

Supplementary Table 1. Differences ( $d$ ; with 95% CI) between lies and truths for unguided judgments using any possible cue (Study2:  $n=30$ ; Study3:  $n=77$ ) or a single cue (verifiability, Study2:  $n=39$ ; Study3:  $n=89$ ; or detailedness, Study2:  $n=39$ ; Study3:  $n=64$ ).

	Single cue judgments			Content coding	
	Unguided (any cue)	Verifiability	Detailedness	Verifiability	Detailedness
Study2	0.39 (0.02; 0.76)	0.81 (0.44; 1.17)	1.06 (0.66; 1.45)	0.69 (0.18; 1.19)	0.77 (0.26; 1.27)
Study3	-0.10 (-0.33; 0.12)	1.23 (0.95; 1.50)	1.15 (0.84; 1.47)		
Stimulus Set 1	0.49 (0.06; 0.91)	1.47 (0.92; 2.02)	1.49 (0.89; 2.06)	1.14 (0.06; 2.19)	0.78 (0.25; 1.79)
Stimulus Set 2	0.09 (-0.27; 0.46)	1.63 (1.12; 2.13)	1.23 (0.69; 1.76)	0.44 (-0.56; 1.42)	0.50 (-0.50; 1.49)
Stimulus Set 3	-0.01 (-0.39; 0.37)	0.82 (0.40; 1.12)	1.64 (1.05; 2.22)	0.55 (-0.46; 1.54)	1.59 (0.43; 2.71)
Stimulus Set 4	-0.42 (-0.81; 0.02)	0.68 (0.30; 1.04)	0.72 (0.30; 1.14)	0.41 (-0.59; 1.39)	0.54 (-0.47; 1.53)

Note. The four bottom rows show the data split per stimulus set (for Study 2 and 3 combined; to restrict loss of power). The last two columns show lie-truth differences obtained for content-coding by trained coders.

