Received: 13 December 2021 | Revised: 23 March 2022 | Accepted: 22 July 2022

DOI: 10.1002/acp.3990

RESEARCH ARTICLE

WILEY

A comparable truth baseline improves truth/lie discrimination

Glynis Bogaard[1] | Madeleine Nußbaum[1] | Laura Sophie Schlaudt[1] | Ewout H. Meijer[1] | Galit Nahari[2] | Aldert Vrij[3]

[1]Department of Clinical Psychological Science, Maastricht University, Maastricht, The Netherlands

[2]Department of Criminology, Bar-Ilan University, Ramat Gan, Israel

[3]Department of Psychology, University of Portsmouth, Portsmouth, UK

Correspondence
Glynis Bogaard, Department of Clinical Psychological Science, Maastricht University, PO Box 616, 6200 MD Maastricht, The Netherlands.
Email: glynis.bogaard@maastrichtuniversity.nl

Funding information
Nederlandse Organisatie voor Wetenschappelijk Onderzoek, Grant/Award Number: VI.Veni.201G.016

Abstract

In a comparable truth baseline (CTB), a knowingly truthful baseline statement is compared to a statement of interest, and deviations in verbal details possibly indicate deceit. In two experiments, we investigated whether a CTB can improve truth/lie discrimination when verbal details are coded by independent raters (Experiment 1) and when judged by naive observers (Experiment 2). In addition, we investigated whether lie tellers would calibrate their lies to match the detailedness of their baseline. Results showed no evidence of calibration. As expected, truths were more detailed than their corresponding baselines, while lies were less detailed. Significant differences emerged for spatial, visual and action details. Experiment 2 did not show that a CTB improved observers' lie detection accuracy. Taken together, our results showed that a deviation in details from a CTB may serve as a helpful aid in lie detection.

KEYWORDS
comparable truth baseline, deception detection, lie detection accuracy, reality monitoring, verbal cues

1 | INTRODUCTION

Lay people and professionals are not particularly good at discerning truth from lies, with on average only around 54% of truths and lies being detected (Bond & DePaulo, 2006). In part, this modest accuracy is due to people relying on non-diagnostic cues to guide their veracity judgements (Akehurst et al., 1996; Levine & Daiku, 2019; Mann et al., 2004). The stereotypical belief, for example, is that non-verbal cues (e.g., gaze aversion, increased body movements) indicate deceit (Strömwall & Granhag, 2003), whereas these cues have been shown to be undiagnostic (DePaulo et al., 2003). In addition, when people do rely on valid cues to deception, their diagnostic utility is not strong. This makes it difficult to be accurate in lie detection even when relying on valid cues (Hartwig & Bond, 2011). Yet, verbal cues, especially the amount of detail, appears to be helpful in detecting deceit (Luke, 2019). Researchers have therefore advocated a refocus, encouraging professionals to listen more carefully to speech instead of relying too much on non-verbal cues (Vrij, 2008b).

Even when relying on more diagnostic verbal cues, one significant obstacle remains, namely that most research evidence relies on group comparison whereas results such as 'lie tellers report fewer details than truth tellers' cannot be easily translated to a single case. For investigators it is often impossible to judge how many details are sufficient to deem a statement truthful (Nahari & Vrij, 2015). There are, for example, large interpersonal differences between interviewees (Bond & DePaulo, 2008). Fantasy prone people, for instance, are better at incorporating details in their stories causing their lies to be judged more believable (Merckelbach, 2004; Schelleman-Offermans & Merckelbach, 2010). Similar findings have been reported for people who are socially skilled (Kashy & DePaulo, 1996; Vrij et al., 2002). People who score high on social adroitness typically provide more details, whether they are true or not (Vrij et al., 2004). More recently,

research has shown that females tend to provide more details than men (Nahari & Pazuelo, 2015).

Because of large individual differences in reporting details, practitioners prefer within-subject methods to make veracity judgements. That is, they prefer to compare the responses from the same person in different parts of the interview. Support for this idea comes from early research showing that being familiar with the normal communicative behaviour of the sender improved observers' lie detection accuracy (Brandt et al., 1980, 1982). In investigative interviewing, however, practitioners and interviewees are rarely acquainted and can only use the time of the interview to familiarise. As a result, interviewers attempt to establish a baseline of an interviewee's truthful communicative behaviour during small talk (Moston & Engelberg, 1993). The behaviour displayed during this small talk is compared to behaviour shown during the investigative phase of the interview, the part for which the veracity is unknown. Any deviation from the baseline may then be seen as an indicator of deceit. Yet, this comparison is problematic as interviewees' responses are expected to differ between these different situations due to factors other than deception. For example, because of differences in stakes between the small talk and investigative part of the interview (Hartwig & Bond, 2014; Ioannou & Hammond, 2015). Unsurprisingly, there is little to no empirical support that small talk baselining is effective in detecting deception (Ewens et al., 2014; Palena et al., 2018).

For a baseline to be of diagnostic value, it should be comparable to the statement of interest, meaning 'that the baseline the investigator uses must be similar in content, context, stakes, and cognitive and emotional involvement to the investigative questions' (Vrij, 2016, p. 1114). These types of statements have therefore been termed comparable truth baseline (CTB). Two recent studies looked at the use of such a CTB. Palena et al. (2018) showed some evidence that incorporating a verbal CTB could increase lie detection accuracy. Truthful and deceptive statements did not differ in detailedness when the small talk was used as a baseline. However, when a CTB baseline was used, the difference in reported spatial details between the second task and CTB differed more for lie tellers than for truth tellers ($d = 1.20$). None of the other details significantly differed between truth tellers and lie tellers. More recently, Bogaard et al. (2022) investigated whether using a truthful baseline could improve truth/lie accuracy. Like Palena et al. (2018), no differences in overall detailedness emerged when comparing deceptive and truthful statements to their baselines. Furthermore, truth/lie discrimination based on verbal differences was comparable with or without using a baseline. In sum, the evidence that a CTB improves truth/lie discrimination is modest at best.

One potential limitation of CTB veracity assessment that can explain the modest accuracy increase is that lie tellers might try to match the detailedness of their deceptive and truthful statements. That is, lie tellers may attempt to make their deceptive statement equally detailed as their truthful baseline. Especially in settings where the veracity of information is unknown yet critical to be determined, lie tellers are likely to be aware that their statements are scrutinised and perhaps compared (Nahari, 2019). Indeed, Verigin et al. (2019) found that lie tellers—whose lies were flanked by truths—included more details in their lies than lie tellers who only told lies. Two more recent studies, however, failed to support calibration. Tomas et al. (2021) asked French undergraduate students to narrate two consecutive statements about their weekend, one truthful and one deceptive. Second narratives were significantly less detailed than first narratives, even more so when these second narratives were deceptive, meaning lie tellers were largely unsuccessful in calibrating their responses. Bogaard et al. (2022) compared the detailedness of target statements that were or were not preceded by a baseline. In line with Tomas et al. (2021), no clear evidence of calibration was found.

Our main aim was to replicate and extend the experiments of Bogaard et al. (2022) but with a different type of statements. Instead of asking participants to provide accounts about negative autobiographical events, participants in the current experiment reported about two crime scenarios they viewed. We aim to investigate to what extent a baseline can improve truth/lie discrimination. However, to this end, we first investigated whether providing a truthful baseline first alters the detailedness of the subsequent target statement for lie tellers as some calibration might occur. If lie tellers are successful in matching their deceptive statements to their baselines, we expect their deceptive statements to be more detailed when they are preceded by a baseline than when they are not (Hypothesis 1). Next, we examined whether there were differences in detail richness between the CTB and target (true and false) statement of the same interviewee. We expected that truth tellers would provide a target statement that is equally or more detailed than their baseline, while lie tellers give a less detailed statement than their baseline (Hypothesis 2).[1] Lastly, we examined whether a CTB would improve truth/lie discrimination as measured by verbal cues. We expected that the difference in verbal cues between truth tellers' and lie tellers' target statements will increase when including the CTB (Hypothesis 3). In Experiment 2, we investigated whether a CTB would improve intuitive judgements of human judges. We expected that observers who received a CTB will achieve higher overall lie detection accuracy rates when judging the target statement than those who do not receive a CTB.

2 | EXPERIMENT 1

2.1 | Methods

2.1.1 | Participants

Based on a G*power analysis (F test; Fixed effects, special, main effects and interactions), with $f = .25$, power of .80 and a standard .05 alpha level, for a 2 (Baseline: CTB vs. no CTB) × 2 (Veracity: lie vs. truth) between-subjects design (see below), a minimum of 128 participants should be included. The total sample included 172 participants (136 female and 36 male participants), ranging in age from 18 to 32 years old ($M = 20.74$, $SD = 2.19$). Participants were undergraduate and graduate students. Participation was rewarded with research participation credits. As an incentive to appear convincing, participants were told that the five most honest appearing interviewees would be rewarded with additional €10 in VVV-vouchers. However, all

participants automatically entered the raffle upon participation and had an equal chance of winning the monetary reward, regardless of their performance. The experiment was approved by the ethical committee of our university and was performed in accordance with the ethical standards of the institution and with the 1964 Helsinki declaration.

2.1.2 | Design

The experiment employed a 2 (Baseline: CTB vs. no CTB) × 2 (Veracity: lie vs. truth) between-subjects design, resulting in four conditions: Baseline Truth ($n = 46$); Baseline Lie ($n = 41$); No Baseline Truth ($n = 45$) or No Baseline Lie ($n = 40$). Participants were randomly assigned to one of these conditions. The total frequency of verbal cues constituted the dependent variable.

2.1.3 | Materials

Stimulus material

Two video sequences, each depicting a mock crime of comparable length and content, were utilised as stimulus material. Both videos show a bike theft committed by two (Video S2) or three (Video S1) individuals. In the first video (Video S1), a man locks his bike and enters a building, which is observed by a woman. After greeting two men, the woman briefly follows the victim to ensure he has entered the building. Shortly after, she returns and, together with the two men she has greeted before, steals the bike. In the second video (Video S2), a woman and a man observe a man locking his bike. The woman follows the victim inside the building. While the woman distracts the bike-owner, the male perpetrator steals the victim's keys and other personal items. When the victim notices the missing belongings, the culprits have already unlocked his bike and fled the scene.

2.1.4 | Procedure

The experiment was administered online via Qualtrics. Prior to the beginning of the experiment, participants were briefly informed about the procedure of the experiment. After signing the informed consent, demographic data was gathered concerning participants' gender, age, and education level. Participants were randomly allocated to the experimental conditions. Depending on the baseline condition, one or two video sequences were presented. For each mock crime video, a time control measurement was utilised to ensure that participants could only continue with the experiment after watching the entire video sequence. Participants in the CTB conditions were first shown Video S1 and asked to provide a written witness statement concerning the content of the presented mock crime. This statement was utilised to establish a CTB. Next, they were shown Video S2 and participants were either instructed to tell the truth or lie about the content of Video S2. Thus, participants in the CTB condition received the instruction to lie or tell the truth only after they provided their baseline but had unlimited time to prepare and write their statement. Participants in the control conditions only viewed Video S2 and provided a truthful or deceptive target statement.

After the statements were provided, to ensure the correct implementation of the veracity instructions, participants expressed the extent to which they reported the truth within their CTB and target statement on a 7-point Likert scale. Moreover, participants indicated how much they agreed with the following statements: 'I found it difficult to give the statement', and 'I was motivated to give the statement' (7-point Likert scales; 1 = strongly disagree; 7 = strongly agree). Finally, participants were fully debriefed concerning the purpose, theoretical background, research questions, and hypotheses of the experiment.

2.1.5 | Reality monitoring scoring

To score the statements' richness in detail, we used the reality monitoring (RM) approach. Like Palena et al. (2018), the frequencies of the RM subcategories listed below were scored. These criteria were selected as they are among the most revealing veracity indicators (DePaulo et al., 2003; Hauch et al., 2017; Masip et al., 2005). A specific detail was only considered once, even if it was mentioned repeatedly. The underlined parts illustrate the details coded:

1. Visual details (e.g., 'A *man* with *red shorts*' [3 details]),
2. Auditory details (e.g., 'The woman *asked for directions*' [1 detail]),
3. Action details (e.g., 'he *opened up* his laptop' [1 detail]),
4. Spatial details (e.g., 'he entered *the kitchen* and put his keys *on the table*' [2 details]), and
5. Temporal details (e.g., 'It happened *after* my lunch break, *in the afternoon*' [2 details]).

Both raters had previously received RM coding training (1.5–2 h), and already had experience with coding statements for previous studies. The rater with the most experience coded all statements while the second rater coded a random selection of 50% of the statements. Both raters were blind to the veracity of the statements. We used the interclass correlation coefficient (ICC) two-way mixed model to calculate the coders' consistency. The average interrater reliability for baseline statements was .92 (lie: ICC = .90; truth: ICC = .93) and for the target statements was .94 (No baseline Lie: ICC = .95; No baseline Truth: ICC = .95, Baseline Lie: ICC = .90; Baseline Truth: ICC = .91). Hence, there was a good to excellent agreement between raters. Separate criteria scores were summed up to get an *RM total score* for the baseline statement, and one for the target statement.

3 | RESULTS

All data were analysed using SPSS Statistics 27. When possible, we reported the effect sizes and accompanying Bayes factors (BFs) using JASP. Bayesian statistics allow to make statements about both H1

TABLE 1 Means, standard deviations and 95% confidence intervals (CIs) of self-reported motivation, perceived task difficulty and self-reported truthfulness

		Motivation			Difficulty			Truthfulness baseline			Truthfulness target		
Baseline	Veracity	*M*	*SD*	95% CI	*M*	*SD*	95% CI	*M*	*SD*	95% CI	*M*	*SD*	95% CI
Yes	TT	5.35	1.12	[5.02; 5.70]	3.14	1.48	[2.67; 3.62]	6.80	0.45	[6.67; 6.94]	6.80	0.40	[6.69; 6,92]
	L	5.37	1.01	[5.05; 5.70]	3.15	1.48	[2.67; 3.61]	6.83	0.38	[6.70; 6.95]	1.59	1.05	[1.24; 1.92]
No	TT	6.02	0.86	[5.76; 6.28]	3.22	1.77	[2.70; 3.75]				6.69	0.63	[6.50; 6.88]
	L	5.12	1.09	[4.77; 5.47]	3.95	2.01	[3.30; 4.59]				2.20	1.40	[1.75; 2.65]

Note: Ratings made on a 7-point Likert Scale (1 = *strongly disagree* to 7 = *strongly agree*).
Abbreviations: L, lie tellers; TT, truth tellers.

and H0 (Jarosz & Wiley, 2014). Because the interaction model in JASP includes both main effects and interaction effects, the interaction model was calculated as (interaction model)/(main factors) (Wagenmakers et al., 2018). Evidence in favour of H_1 is indicated as BF_{10}, whereas evidence in favour of H_0 is indicated by BF_{01}. We used the following classification scheme: BF > 100 extreme evidence, 30–100 = very strong evidence, 10–30 = strong evidence, 3–10 = moderate evidence, 1–3 = anecdotal evidence and 1 = no evidence for H_1 (for more information on interpretation of BFs see Lee & Wagenmakers, 2013).

3.1 | Motivation and task difficulty

We performed 2 two-way analysis of variances (ANOVAs) with veracity and baseline as between-subjects factors and self-reported motivation and perceived task difficulty as the dependent variables. Table 1 shows that, overall, participants appeared to be motivated to perform well on the task ($M = 5.50$, $SD = 1.07$, 95% CI [5.32; 5.63]). Results did show a significant interaction effect, $F(1, 172) = 8.54$, $p = .004$, $\eta_p^2 = .05$, $BF_{10} = 10.37$. Follow up simple effects showed that truth tellers were more motivated that lie tellers in the No Baseline condition, $F(1, 83) = 17.83$, $p < .001$, $\eta_p^2 = .18$, $BF_{10} > 100$. No significant differences emerged in the Baseline condition, $F(1, 85) = .007$, $p = .93$, $\eta_p^2 < .001$ $BF_{01} = 4.44$. Participants did not find the task very difficult ($M = 3.33$, $SD = 1.71$, 95% CI [3.10; 3.61]) and no significant differences emerged between groups (all *F*'s < 3.31, all *p*'s > .07). Yet, support for the null hypothesis was anecdotal (Veracity: $BF_{10} = 1.51$; Baseline: $BF_{10} = 2.14$; interaction: $BF_{10} = 2.09$). *Ms*, *SDs* and CIs for the separate conditions are presented in Table 1.

3.2 | Manipulation check

Participants who provided a baseline reported to have told the truth in their baseline. Baseline truthfulness did not significantly differ between the veracity groups [$t(84) = -.23$, $p = .82$, $d = .05$, $BF_{01} = 4.30$] showing participants adhered to our instructions for providing their baseline. See Table 1 for *Ms*, *SDs* and CIs.

To investigate the self-reported truthfulness of the target statement, we performed a two-way ANOVA with Veracity and Baseline as between-subjects factors and self-reported truthfulness of the target statement as dependent variable. Results showed no significant main effect of Baseline, $F(1, 168) = 3.08$, $p = .08$, $\eta_p^2 = 02$, $BF_{01} = 5.18$, whereas the Veracity main effect, $F(1, 168) = 1167.80$, $p < .001$, $\eta_p^2 = .87$, $BF_{10} > 100$, and Baseline × Veracity interaction effect, $F(1, 168) = 6.61$, $p = .01$, $\eta_p^2 = .04$, $BF_{10} = 4.00$, were significant. Of these two significant effects, the interaction effect is the most informative. A subsequent test of simple effects, separated per level of baseline, showed that the differences between lie tellers and truth tellers in both Baseline conditions were large. However, the effect in the Baseline condition, $F(1, 85) = 980.45$, $p < .001$, $\eta_p^2 = .92$, $BF_{10} > 100$ was slightly larger than in the No Baseline condition, $F(1, 83) = 376.60$, $p < .001$, $\eta_p^2 = .82$, $BF_{10} > 100$. As instructed, in both Baseline conditions, lie tellers indicated that they were significantly less truthful than truth tellers. These results showed that our manipulation was successful.

3.3 | Word count

Investigation of *the baseline statements* showed that the provided baselines were on average 134 words long ($SD = 51.68$) and that length did not significantly differ between truth tellers ($M = 125.13$, $SD = 53.11$, 95% CI [109.36; 140.90]) and lie tellers ($M = 143.88$, $SD = 48.78$, 95% CI [128.48; 159.27]), $F(1, 85) = 2.91$, $p = .09$, $\eta_p^2 = .03$. Note that support for the null hypothesis was anecdotal $BF_{01} = 1.25$.

On average, the *target statements* were 151 words long ($SD = 72.52$). A two-way ANOVA with Veracity and Baseline as between-subject factors showed that truth tellers provided longer target statements ($M = 165.33$, $SD = 73.92$, 95% CI [150.81; 180.12]) than lie tellers ($M = 133.98$, $SD = 67.58$, 95% CI [118.41; 149.47]), $F(1, 168) = 8.49$, $p = .004$, $\eta_p^2 = .05$, $BF_{10} = 7.50$. Results showed no significant difference in the length of the target statements between participants who provided a baseline ($M = 145.07$, $SD = 61.97$, 95% CI [129.59; 59.61]) and those who did not ($M = 156.19$, $SD = 81.93$, 95% CI [139.63; 170.01]) $F(1, 168) = 0.89$, $p = .35$, $\eta_p^2 = .01$, $BF_{01} = 3.80$. The interaction effect was not significant either, $F(1, 168) = 1.95$, $p = .16$, $\eta_p^2 = .01$, but evidence for the null hypothesis was anecdotal $BF_{10} = 1.91$.

3.4 | Hypotheses testing

3.4.1 | Calibration of deceptive statements to match the baseline

To investigate whether lie tellers calibrate their deceptive statement to match their baseline, we tested the effect of Baseline for lie tellers. That is, we conducted an ANOVA with Baseline as the between-subjects factor and the total number of RM details included in the target statement of lie tellers as the dependent variable. Evidence of calibration is found if lie tellers in the baseline group include significantly more details in their target statement than in the no baseline group.

For lie tellers, results showed no significant difference in the detailedness of deceptive target statements when they were ($M = 49.61$, $SD = 20.62$) or were not preceded by a Baseline ($M = 43.03$, $SD = 23.77$), $F(1, 79) = 1.78$, $p = .19$, $\eta_p^2 = .02$, $BF_{01} = 2.01$. Note that based on the Bayesian Factors support for these findings was anecdotal. Nonetheless, at the very least we can conclude that there was no strong evidence for Hypothesis 1, namely that lie tellers matched their deceptive target to their truthful baseline.

3.4.2 | Do truths and lies differ in their detail richness to the CTB?

For the subsequent analyses, we included only participants from the baseline group. The RM baseline score did not differ between truth tellers ($M = 48.13$, $SD = 18.41$, 95% CI [42.66; 53.60]) and lie tellers ($M = 55.02$, $SD = 18.47$, 95% CI [49.20; 60.85]), $F(1, 85) = 3.03$, $p = .08$. A lack of pre-existing RM differences between truth tellers and lie tellers in their baseline allows us to reliably investigate the RM difference score as an indicator of deception.

We compared the difference in detail richness between CTBs and their corresponding target statements by examining the effect of Veracity (lie vs. truth) on the *directional RM difference scores*, calculated as follows [Baseline RM score − Target RM score]. A negative score means that the target statement is more detailed than the baseline statement, while a positive score indicates the opposite. We conducted an ANOVA with Veracity as factor and *the directional RM difference score* as the dependent variable. Results were in the expected direction and revealed a significant Veracity effect, $F(1, 85) = 15.41$, $p < .001$, $d = 0.84$, $BF_{10} > 100$. Truth tellers provided a more detailed target statement than their baseline statement ($M = -7.72$, $SD = 16.93$, 95% CI [−12.75; −2.69]), while lie tellers provided a less detailed target statement ($M = 5.41$, $SD = 13.89$, 95% CI [1.03; 9.80]).[2] Next, we tested whether these scores significantly deviated from zero using one-sample *t* tests. This was the case for both truth tellers [$t(45) = -3.09$, $p = .003$, $d = .46$, $BF_{10} = 9.86$] and lie tellers [$t(40) = 2.50$, $p = .017$, $d = .40$, $BF_{10} = 2.61$]. Thus, moderate support for Hypothesis 2 was found for truth tellers, while anecdotal support was found for lie tellers.

Exploratory analyses

We further explored whether the RM difference score for the individual RM criteria (instead of the RM total score) would differ between truth tellers and lie tellers. Results of the multivariate analysis of variance showed a significant multivariate Veracity effect, $F(5, 81) = 4.06$, $p = .002$, $\eta_p^2 = .20$. At the univariate level, results showed that for truth tellers the RM difference score was significantly lower for spatial, visual and action details than for lie tellers, see Table 2.

3.4.3 | Does a CTB improve truth/lie discrimination?

To test whether the inclusion of a CTB enhanced the truth/lie discrimination, we tested whether the effect size of the difference in verbal cues between lie tellers and truth-tellers increased when a CTB was included as a covariate. An ANOVA with Veracity as factor and RM total target scores as dependent variable yielded no significant effect, $F(1, 85) = 2.03$, $p = .16$, $\eta_p^2 = .02$, $d = .31$. An analysis of covariance with Veracity as factor, CTB as covariate and RM total target scores as dependent variable revealed a significant Veracity effect, $F(1, 83) = 4.16$, $p = .04$, $\eta_p^2 = .05$, $d = .41$. Thus, after taking into consideration the RM baseline score, truths were significantly more detailed than lies, supporting Hypothesis 3.

4 | DISCUSSION

We found no clear evidence that lie tellers matched the detailedness of their target statements to their truthful baselines. Next, we examined whether using a truthful baseline statement as a within-subject

TABLE 2 *M*s, *SD*s, 95% CIs and significance for the separate RM criteria for the directional RM difference score

	Truth tellers			Lie tellers					
Criteria	***M***	***SD***	**95% CI**	***M***	***SD***	**95% CI**	***F***	***p***	η_p^2
Spatial	−5.10	4.96	[−6.56; −3.65]	−2.63	4.96	[−4.17; −1.09]	5.39	.02	.06
Temporal	−1.43	2.39	[−2.20; −0.66]	−0.94	2.39	[−1.71; −0.91]	0.90	.34	.01
Auditory	−1.76	1.19	[−2.32; −1.20]	−1.63	2.46	[−2.22; −1.04]	0.09	.76	<.01
Visual	4.97	10.80	[1.97; 8.95]	12.14	10.06	[9.17; 15.66]	10.94	<.01	.11
Action	−4.39	2.95	[−5.38; −3.40]	−1.83	3.82	[−2.88; −0.77]	12.38	<.01	.13

Note: Criteria indicated in bold show a significant difference between truth tellers and lie tellers ($p < .02$).

comparison would improve lie detection performance. Results revealed that for truth tellers, target statements were more detailed than their baselines, while for lie tellers, baseline statements were more detailed than the corresponding target statements. Furthermore, our results showed that after taking into consideration the RM baseline score, true target statements were more detailed than deceptive statements.

In line with Tomas et al. (2021) and Bogaard et al. (2022), we found no clear evidence of calibration. Tomas et al. (2021) also reported an overall decline in detailedness, but more so for lies than truths. In our experiment, we found no influence of baseline on the length of subsequent target statements (true or false), and evidence in support of this finding was moderate. Overall, these findings show that lie tellers seem to be largely unable to match the verbal content of their deceptive statement to their truthful baseline.

Results further revealed that for truth tellers target statements were more detailed than their baselines, while for lie tellers, baseline statements were more detailed than the corresponding target statements. Exploratory analyses of the RM difference score revealed that truth tellers and lie tellers especially differed in the number spatial, visual and action details they included in their statements. Palena et al. (2018) found that only spatial details resulted in detectable differences between lie tellers and truth tellers. The diverging results between our experiment and Palena et al. (2018) might be explained by the differing methods employed to compare the detailedness of lie tellers' and truth tellers' statements. While Palena et al. (2018) compared the detailedness of true and false accounts by means of the absolute verbal difference between baseline and target statements, we examined the presence of a directional difference. Indeed, when applying the absolute RM difference score on our data, results show no significant difference between truths and lies (see endnote 2). Bogaard et al. (2022) also used the directional difference approach and found a difference in auditory and temporal details, but not in total details. The fact that in their study other types of details were related to veracity, may be explained by a difference in the statement type used. The current study included two mock crimes, while participants in Bogaard et al. (2022) reported about negative autobiographical events. Taken together, these results suggest that the increase or decrease in total details from a comparable baseline statement, rather than the absolute difference in RM details, may serve as a helpful aid in lie detection.

Lastly, we examined whether using a CTB can improve truth/lie discrimination. Our results showed that after taking into consideration the RM baseline score, true target statements were more detailed than deceptive statements. Like Bogaard et al. (2022), the observed effect was small to moderate ($d = .41$). In contrast, other content-based lie detection tools, such as criteria-based content analysis (CBCA), have been shown to achieve moderate effects ($dw = 0.55$; Amado et al., 2016; $g = 0.58$ after controlling for publication bias; Oberlader et al., 2016). Thus, established techniques such as CBCA seem to have somewhat higher effect sizes than CTBs in terms of deception detection.

Taken together, our results show that CTBs can improve truth/lie discrimination. Truth tellers tend to include more details in their target statement, while lie tellers include less as compared to their baseline.

5 | EXPERIMENT 2

Results of Experiment 1 suggest that the use of a CTB can improve truth/lie discrimination as measured by the increase in the RM difference score. In investigative interviewing, however, conclusions about the interviewee's veracity often reflect a subjective judgement rather than a formal analysis of content criteria. The objective of Experiment 2 was to examine whether CTB can improve laypeople's veracity judgements.

Few studies to date have tested whether access to a CTB can improve observers' judgements. In Caso, Palena, Vrij, and Gnisci (2019) laypeople were presented with videotaped interviews divided into two parts: a baseline (small talk vs. CTB) and a target part. They were also informed that the baseline statement was always truthful, whereas the veracity of the target statement was either true or false. The task was to judge the veracity of the target statements using the truthful baseline as a decision aid. In the CTB condition, observers' total accuracy rates scored significantly above chance (56.5%), whereas accuracy rates of observers in the small talk condition did not reach such levels (47.4%). In Caso, Palena, Carlessi, and Vrij (2019), a similar procedure was used, but with police officers instead of laypeople. Their results showed that the presentation of a CTB significantly improved police officers' lie detection accuracy (54.7%) compared to that of police officers receiving no baseline (39.2%). No differences were found between groups in truth and total accuracy.

Other studies, however, did not show evidence that providing observers with a truthful baseline improved their veracity judgements. For example, Verigin et al. (2020) let participants read an alibi statement in which suspects either lied or told the truth about a critical 2-h period. The remainder of the statement was always truthful. Half of the participants was told to use this known truthful information to make a veracity judgement about the critical 2-h period. Results showed no evidence that such a within-statement baseline comparison improved deception detection. In line with these findings, Bogaard et al. (2022) also failed to show that providing a known truthful baseline improved observers' veracity judgement. Contrary to their hypothesis, observers that used a baseline had a lower overall accuracy than observers who did not (45% vs. 56% and 52%).

In Experiment 2, those receiving a CTB were instructed to use this statement to inform their veracity judgement regarding a target statement from the same person. Based on the contradicting findings, we feel it is better not to make precise predictions. The aim of Experiment 2 was to examine to what extent providing a CTB would influence observers' total accuracy, lie accuracy and truth accuracy.

6 | METHODS

6.1 | Participants

An a priori power analysis using G*Power (*F* tests; repeated measures, within-between interaction) with a small effect size ($f = 0.20$), an alpha of .05 and a power of .80 indicated that a minimum sample of

66 was needed. A total of 224 participants completed the online survey (206 women and 16 men, two preferred not to indicate their gender) with an age range from 18 to 63 years ($M_{age} = 25.39$, $SD = 6.53$). Participants were recruited through an online research participation platform and advertisements on social media (e.g., Facebook). Undergraduates were compensated with partial course credit for their participation. The experiment was approved by the ethical committee of our university and was performed in accordance with the ethical standards of the institution and with the 1964 Helsinki declaration.

6.2 | Design

The present experiment utilised a 3 (Baseline Type: No baseline vs. Target vs. Baseline-Target) × 2 (Veracity: Truth vs. Lie) mixed design with Baseline Type as a between subject factor and Veracity as a within subject factor. The dependent variables were observers' total, truth and lie accuracy rates as well as self-reported believability.

6.3 | Materials

The statements that were obtained in Experiment 1 were used for this experiment. Statements about the first video clip served to establish a truthful baseline and statements about the second video clip were presented as the target statements. Four statements (two truthful and two deceptive) were randomly presented to participants based on the experimental condition they were randomly allocated to.

6.4 | Procedure

The experiment was administered online via Qualtrics. After giving informed consent, participants were randomly allocated to one of the three experimental conditions. The No baseline condition ($n = 78$) received two truthful and two deceptive target statements randomly drawn from a pool of 17 statements of mock witnesses that did not give a baseline statement in Experiment 1. Participants in the Target condition ($n = 74$) received two truthful and two deceptive target statements randomly drawn from a pool of statements of mock witnesses that did give a baseline statement in Experiment 1. However, participants in this condition received the target statements without their corresponding baseline statement. Finally, participants in the Baseline-Target condition ($n = 72$) randomly received two truthful and two deceptive target statements from the same pool of statements as the Target condition, the difference being that this time both baseline and target statement of the same person were presented.

Participants were briefed about the aim of the experiment and told that they were about to receive several statements of mock witnesses, their task being to determine the veracity of these statements. They were also informed that for each statement they would only be able to proceed to the next statement after a timer of 1 min had passed to make sure they would carefully read each statement. Participants in the No Baseline and Target conditions were not given further instructions. Participants in the Baseline-Target condition were told that they would be presented with two statements from the same person, the first being truthful while the veracity of the second statement was unknown, hence doubtful. They were instructed to use the first statement to inform their judgement regarding the second statement. Indications about which speech patterns might indicate deceit were not provided.

Participants in all conditions received four statements only: two truthful and two deceptive ones. This relatively small number of statements was chosen to avoid that participants would recognise a pattern in the content of the statements, as all truthful target statements were about the same mock crime. Presenting more statements could have potentially increased the risk that participants would base their judgement on content rather than looking at differences in speech patterns.

For each statement, participants were asked to answer the following questions: (1) 'How believable do you judge this statement?' (7-point Likert scale; 1 = extremely unbelievable; 7 = extremely believable) and (2) 'If you have to choose, you find this statement to be more…' with the dichotomous options 'True' and 'False'. Lastly, participants were asked whether they were motivated to perform well in the experiment, whether they thought judging the veracity of the statements was easy for them, whether they thought they did well in judging the veracity of the statements and whether they usually get away with lies (7-point Likert scale; 1 = strongly disagree; 7 = strongly agree). In the end, participants were thanked for their participation and fully debriefed about the aim of the experiment.

7 | RESULTS

7.1 | Statistical analyses

All data were analysed using SPSS Statistics 27. We also used JASP to report Bayesian Factors. To investigate whether including a baseline statement improved participants' truth and lie accuracy, we first calculated the number of correct judgements (coded as 0 = incorrect, 1 = correct) for true and false statements separately with a maximum score of two for each Baseline Type. Scores were averaged for truth and lies separately, resulting in two measures per participant. For overall accuracy, scores on all four statements were averaged. For the main hypotheses, we conducted mixed ANOVAs with Baseline Type as between factor and Veracity as within factor. In addition, we performed the same analyses as Bayes ANOVAs using JASP.

7.2 | Motivation and task difficulty

First, through four ANOVAs with Baseline Type as factor we checked whether any of the groups differed in terms of motivation to perform well, perceived task difficulty, subjective performance evaluation and how well they usually get away with lies. None of the effects were significant, all Fs < 1.47, all p's > .22. Support for the null hypothesis

was moderate for the perceived task difficulty ($BF_{01} = 5.87$) and strong for all other variables ($BF_{01} > 10$). On average, all participants were very motivated ($M = 5.82$, $SD = 0.99$). For the other question participants' average answers varied around the neutral option: difficulty ($M = 3.60$, $SD = 1.57$), performed well ($M = 4.13$, $SD = 1.17$), get away with lies ($M = 4.30$, $SD = 1.53$), see Table 3.

7.3 | Hypotheses testing

7.3.1 | Accuracy (binary choice)

Total accuracy

A one-way (Bayesian) ANOVA with overall accuracy (proportion correct) as dependent variable and Baseline Type as factor revealed that there were statistical differences between the Baseline Type conditions, $F(2, 221) = 5.05$, $p = .007$, $\eta_p^2 = .04$, $BF_{10} = 4.04$. Post hoc tests using a Bonferroni correction showed that participants in the Target condition ($M = 0.50$, $SD = 0.26$), $p = .01$, $d = 0.46$, *95%* CI_{diff} [0.02, 0.22], $BF_{10,U} = 5.45$ and Baseline-Target condition ($M = 0.49$, $SD = 0.24$), $p = .03$, $d = 0.42$, *95%* CI_{diff} [0.008, 0.22], $BF_{10,U} = 3.92$ were overall statistically more accurate than participants in the No Baseline condition ($M = 0.38$, $SD = 0.28$). There were no differences between the Target and the Baseline-Target condition, $p = 1.00$, $d = 0.04$, *95%* CI_{diff} [−0.95, 0.12], $BF_{01,U} = 5.46$. Support for these findings was moderate based on the accompanying BFs.

Truth and lie accuracy

To investigate whether there were any differences in terms of accuracy between the groups for the truthful and deceptive statements separately, we performed a mixed (Bayesian) ANOVA with Baseline type (No baseline, Target, Baseline-Target) as between-subject factor and Veracity (truth vs. lie) as within-subject factor. There was no significant main effect of Veracity, $F(1, 221) = 2.43$, $p = .12$, $\eta_p^2 = .011$, $BF_{01} = 2.41$, but there was a significant main effect of Baseline Type, $F(2, 221) = 5.05$, $p = .007$, $\eta_p^2 = .04$, $BF_{10} = 2.21$. There was also a significant interaction effect between Veracity and Baseline type, $F(2, 221) = 6.87$, $p = .001$, $\eta_p^2 = .06$. Accordingly, the accompanying BF suggests that the model is more than 40 times more likely under the interaction effect than under the two main effects ($BF_{10} = 40.64$).

To examine the interaction effect, we performed pairwise comparisons for each condition and separated for truthful and deceptive statements. Results revealed that for truthful statements, there were no significant differences between conditions (all p's > .23). Support for the null hypothesis between the Target and Baseline-Target condition ($BF_{01,U} = 1.38$) and the No Baseline and the Target condition ($BF_{01,U} = 2.80$) was anecdotal, while support for the null hypothesis between the No Baseline and Baseline-Target conditions was moderate ($BF_{01,U} = 5.01$). For the deceptive statements, results showed that the Target condition and the Baseline-Target condition achieved significantly higher accuracy rates than the No Baseline condition, $p = .009$, $d = 0.48$, *95%* CI_{diff} [0.03, 0.29], $BF_{10,U} = 8.24$ and $p < .001$, $d = 0.75$, *95%* CI_{diff} [0.12, 0.39], $BF_{10,U} > 100$, respectively. There were no significant differences in terms of accuracy between the Target and the Baseline-Target condition, $p = .34$, $d = 0.26$, *95%* CI_{diff} [−0.22, 0.05], but support for the null hypothesis was anecdotal $BF_{01,U} = 1.76$. These findings combined mean that accuracy rates in the Target and Baseline-Target conditions were somewhat superior to those in the No Baseline condition.

Signal detection analyses

Given that we found some differences between conditions for overall accuracy and lie accuracy, we tested whether participants in a specific condition might have shown a response bias. To this end, we calculated *c* for each condition (see Stanislaw & Todorov, 1999). Similar analyses have been reported by Caso, Palena, Carlessi, and Vrij (2019) to examine response bias in veracity judgements. The following interpretation is used for *c*: if the *c* value is greater than 0, there is a bias toward responding 'no', in this case truth; if the *c* value is lower than 0, there is a bias toward responding 'yes', in this case lie. We conducted a one-sample *t* test per condition with c as the dependent variable and 0 as the test score. For the No Baseline condition ($M = .59$, $SD = .62$) participants favoured a 'truth' response, $t(77) = 8.47$, $p < .001$, $d = .62$, $BF_{10} > 100$, which was also the case for the Target condition ($M = .49$, $SD = .76$), $t(73) = 5.53$, $p < .001$, $d = .77$, $BF_{10} > 100$. The Baseline-Target condition ($M = .13$, $SD = .58$) did not show a clear response bias $t(71) = 2.03$, $p = .05$, $d = .58$. However, support for the null hypothesis is anecdotal $BF_{01} = 1.13$. Lastly, we examined whether *d′* scores—a measure of sensitivity (Stanislaw & Todorov, 1999)—differed from chance level responding, indicated by a score of 0. Significant deviations indicate either worse than (<0) or better than (>0) chance performance. Participants in the No Baseline condition ($M = -.68$, $SD = .1.64$) performed significantly worse than chance, $t(77) = -3.66$, $p < .001$, $BF_{10} = 49.44$, while the Target ($M = .11$, $SD = 1.53$) and Baseline target conditions ($M = .02$, $SD = 1.29$), $t(73) = .63$, $p = .52$, $BF_{01} = 6.44$ and $t(72) = .11$, $p = .91$, $BF_{10} = 7.67$ did not differ from chance level performance.

TABLE 3 *M*s, *SD*s and 95% CIs of self-reported motivation, easiness of task, subjective task performance and ability to get away with lies

	Motivation			Difficulty			Performance			Get away		
Baseline type	***M***	***SD***	**95% CI**	***M***	***SD***	**95% CI**	***M***	***SD***	**95% CI**	***M***	***SD***	**95% CI**
No baseline	5.86	1.18	[5.59; 6.13]	3.71	1.46	[3.38; 4.03]	4.21	1.22	[3.93; 4.48]	4.36	1.64	[3.99; 4.73]
Target	5.88	0.79	[5.69; 6.06]	3.74	1.56	[3.38; 4.11]	4.09	1.20	[3.82; 4.37]	4.42	1.29	[4.12; 4.72]
Baseline-target	5.72	0.97	[5.49; 5.95]	3.35	1.61	[2.97; 3.73]	4.10	1.10	[3.84; 4.36]	4.11	1.66	[3.72; 4.50]

Note: Ratings made on a 7-point Likert Scale (1 = *strongly disagree* to 7 = *strongly agree*).

TABLE 4 Overview of the percentage of correct judgements (binary judgement) and the average believability score (7-point scale) for each condition

	Truths				Lies				Total
		Believability				Believability			
Baseline type	% Correct	*M*	*SD*	95% CI	% Correct	*M*	*SD*	95% CI	% Correct
No baseline	47	3.84	1.41	[3.51; 4.16]	29	4.86	1.24	[4.57; 5.15]	38
Target	55	4.13	1.58	[3.80; 4.47]	46	4.20	1.41	[3.91; 4.50]	50
Baseline-target	44	3.75	1.33	[3.41; 4.09]	55	3.88	1.24	[3.5; 4.18]	49

Note: All answers were provided on a 7-point Likert scale (1 = not at all, 7 = very much).

7.4 | Believability (7-point Likert scale)

A mixed (Bayesian) ANOVA with Baseline Type as between-subject factor, Veracity as within-subject factor and believability as dependent variable revealed a significant main effect of Veracity, $F(1, 221) = 10.39$, $p = <.001$, $\eta_p^2 = .05$, $BF_{10} = 25.39$, a significant main effect of Baseline Type, $F(2, 221) = 5.43$, $p = .005$, $\eta_p^2 = .05$, $BF_{10} = 2.57$, and a significant interaction effect between Veracity and Baseline Type, $F(2, 221) = 6.07$, $p = .003$, $\eta_p^2 = .05$, $BF_{10} = 16.50$. The accompanying BFs imply that the model is almost 17 times more likely under the interaction effect than under the two main effects. Analysis of the simple effects using a Bonferroni correction showed that for the truthful statements, there were no significant differences in believability ratings between the conditions (all *p*'s > .57). Support for the null hypothesis between the Target and Baseline-Target condition ($BF_{01,U} = 1.78$) and the Target and No Baseline condition ($BF_{01,U} = 2.90$) was anecdotal. Support for the null hypothesis between the No Baseline and Baseline-Target conditions was moderate ($BF_{01,U} = 5.30$). For the deceptive statements, however, participants in the No Baseline condition believed the statements significantly more often to be true than the Target condition, $p = .006$, *95%* CI_{diff} [.14, 1.16], $d = 0.50$, $BF_{10,U} = 11.58$, and Baseline-Target condition, $p < .001$, *95%* CI_{diff} [.46, 1.49], $d = 0.71$, $BF_{10,U} > 100$ (Table 4). Again, there were no significant differences between the target and Baseline-Target condition, $p = .441$, 95% CI_{diff} [−.20, .83], $d = 0.24$, yet support for the null hypothesis was anecdotal $BF_{01,U} = 2.12$. This pattern of results resembled the accuracy score results.

8 | DISCUSSION

Results of Experiment 2 showed that both the Target and the Baseline-Target conditions significantly outperformed the No Baseline condition in terms of total and lie accuracy. Support for most of these findings was moderate, except for the difference in lie accuracy between the Baseline-Target and No Baseline condition. Here we found strong support that the Baseline-Target condition outperformed the No Baseline condition. Results showed no clear evidence that differences in accuracy existed between Target and Baseline-Target condition for any type of statement. Comparable results were found for believability (7-point Likert scale). Given both the Target and Baseline-Target conditions outperformed the No Baseline condition, familiarity with a person's truthful verbal behaviour by presenting a CTB did not aid observers in their veracity judgement of a second statement.

In line with Bogaard et al. (2022), we found that observers who were provided with a baseline did not outperform observers who did not. Results showed that both the Baseline and the Target-Baseline conditions were overall more accurate than the No Baseline group, which indicates that the mere act of asking people for a baseline statement helped observers detecting deception in target statements. Why this is the case is unclear. Experiment 1 did not show evidence that providing a baseline altered the detailedness of deceptive statements. Furthermore, target statements were of a comparable length, regardless of whether participants provided a baseline. Therefore, it is more likely that observers relied on other cues than detailedness. However, given that we did not ask participants which cues they based their judgements on, it is unclear what these could be. Future studies should investigate this.

In line with the findings of Caso, Palena, Carlessi, and Vrij (2019), investigation of participants' response bias showed strong evidence that the conditions in which no baseline was presented displayed a so-called truth-bias. That is the tendency of naïve observers to believe that a speaker is telling the truth independent of actual honesty (Bond & DePaulo, 2006). In everyday life, this presumption of honesty enables efficient communication and often leads to a correct belief, as we are more often confronted with truthful rather than deceptive statements (Levine, 2014). However, no clear evidence of a response bias was detected when a baseline was presented. Being presented with a baseline statement seemed to make people more critical about their veracity decision, but it did not lead to a better-than-chance accuracy. These latter findings are also in line with Caso, Palena, Carlessi, and Vrij (2019).

9 | GENERAL DISCUSSION

Experiment 1 showed no evidence that deceptive individuals were able to manipulate their language use to imitate their baseline language. Furthermore, our finding of a significant directional difference supported the use of a CTB as a within subject measure.

So, this method seems to have potential to be used in lie detection, yet our results also show baselining needs further refinement before it can be implemented in real life settings. A CTB can provide investigators with an idea of how detailed a truthful statement from a specific person is and compare this detail richness to a doubtful statement. However, this still does not allow practitioners to use a simple decision rule 'when the CTB is more detailed than the doubtful statement, the doubtful statement is false', which would be the ultimate goal of baselining. Applying this decision rule to our data resulted in a 61% accuracy for deceptive statements and 65% accuracy for the truthful statements, please see Appendix for the analyses. Truth and lie accuracy using baselines is somewhat lower than the average accuracy of CBCA and RM (Amado et al., 2016; Masip et al., 2005; Oberlader et al., 2016), but these numbers seem promising.

Furthermore, previous experiments, including the current one, have mostly generated CTBs from initial, separate portions of an interview and compared these to a statement of interest. In practice, however, it appears much more desirable to be able to draw a truthful baseline from parts of a single statement rather than having to generate a baseline statement in a separate part of the interview that precedes the investigative part. This is what Verigin et al. (2020) examined. Even though their results showed that instructing participants to make this within-statement baseline comparison did not improve the accuracy of their credibility assessments, studies like this invite future research to further investigate the potential of verbal baselining, as the question how to establish a verbal baseline that is most suitable for lie detection purposes still prevails.

In addition, research suggests that more active interviewing approaches, like the reality interview (RI; Colwell et al., 2002), have the potential to enhance verbal differences between truths and lies (Vrij, 2014). The idea behind these so-called cognitive approaches is to magnify the difference in cognitive load for lie tellers and truth tellers, therein making the task for lie tellers even more cognitively demanding (Vrij et al., 2017). Bogaard et al. (2019), for instance, showed that using the RI improved the accuracy of both CBCA (Steller & Köhnken, 1989) and RM (Johnson & Raye, 1981). Theoretically, using baselining in combination with an active interviewing approach such as the RI could result in larger differences between baseline and target statements, making it easier for observers to catch a lie. Future studies should investigate the potential of combining cognitive approaches and within-subject verbal lie detection techniques like CTB to facilitate credibility assessment.

We found no evidence that including a baseline statement improves lay people's lie detection accuracy (Experiment 2). Although the Baseline-Target condition outperformed the No Baseline condition for total and lie accuracy, this condition did not perform better than the Target condition. Furthermore, accuracy rates were overall low and comparable to other studies, showing truth and lie accuracy hoover around chance level (Bond & DePaulo, 2006, 2008). Thus, while CTBs may be promising as a verbal lie detection tool, including a baseline statement does not seem to improve lay people's intuitive judgements.

The results of the present experiments should be considered in light of some limitations. First, the experiments were conducted online, so there was no control by the experimenter that participants fully watched/read the videos/statements. We aimed to address this issue by including time-control measures to ensure participants watched/read the entire statement before they could continue to subsequent questions. However, we did not include additional attention check questions. Second, there were no possible impending negative consequences for participants, as there would be in real-life investigative settings. We aimed at raising the stakes of the experiments by offering a reward for the most convincing interviewees. Still, the stakes of our experiment were likely to be comparatively low compared to an authentic criminal investigation. Furthermore, participants provided written instead of oral testimonies, which gave them the opportunity to rewrite their statements to appear more credible. In addition, participants wrote about what they witnessed rather than what they experienced. These limitations might have influenced their verbal behaviour. Our sample that provided the statements were primarily students and this might have influenced the content of the collected statements. To minimise this influence, we controlled for the events they had to report about, yet statements were overall short. This could have made it difficult for observers to properly use the baseline statements to their advantage.

The current experiments examined the extent to which a baseline can increase truth/lie discrimination. Results seemed promising when verbal cues were taken into consideration (Experiment 1). However, presenting a truthful baseline to naïve observers did not improve their veracity judgement (Experiment 2). Hence, advising the use of baselines in practice is premature. Future studies should reproduce the directional difference between baseline and target statements and whether this difference could serve as a decision rule. Furthermore, in addition to the limitations addressed above, what constitutes as a comparable baseline to be used in practice should be examined further. For example, how comparable should the content of two statements be, to apply baselining?

In conclusion, CTBs may represent a promising addition to the existing lie detection tools that allows for a within-subject comparison. Whilst Experiment 2 failed to support the notion that including a baseline statement would improve lay people's accuracy rates, Experiment 1 showed support for the usefulness of CTBs as a verbal lie detection method. Lie tellers provided less detailed target statements compared to their baselines, while truth tellers tended to include more details in their target than their baselines.

ACKNOWLEDGEMENTS

This publication is part of the project 'Outsmarting liars' with project number VI.Veni.201G.016 of the research programme Veni which is financed by the Dutch Research Council (NWO).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of both experiments are available at the Open Science Framework (see https://osf.io/6reay/).

ORCID

Glynis Bogaard https://orcid.org/0000-0001-6795-9433

Aldert Vrij https://orcid.org/0000-0001-8647-7763

ENDNOTES

[1] Note that this hypothesis deviates somewhat from our pre-registration that states 'We expected that there will be more similarity in verbal cues between truth tellers' CTB and target statement than between lie tellers' CTB and target statement'. This hypothesis was based on the findings of Palena et al. (2018). However, we decided to finetune this hypothesis based on more recent findings (Bogaard et al., 2022). See endnote 2 for results of the pre-registered analysis.

[2] When applying the pre-registered analysis to our data, that is the absolute RM difference score, results showed no significant difference between truths ($M = 13.54$, $SD = 12.65$) and lies ($M = 11.36$, $SD = 9.52$), $F(1, 85) = .81$, $p = .37$, $d = .19$.

REFERENCES

Akehurst, L., Kohnken, G., Vrij, A., & Bull, R. (1996). Lay persons' and police officers' beliefs regarding deceptive behaviour. *Applied Cognitive Psychology*, *10*(6), 461–471. https://doi.org/10.1002/(Sici)1099-0720(199612)10:6<461::Aid-Acp413>3.0.Co;2-2

Amado, B. G., Arce, R., Farina, F., & Vilarino, M. (2016). Criteria-based content analysis (CBCA) reality criteria in adults: A meta-analytic review. *International Journal of Clinical and Health Psychology*, *16*(2), 201–210. https://doi.org/10.1016/j.ijchp.2016.01.002

Bogaard, G., Colwell, K., & Crans, S. (2019). Using the reality interview improves the accuracy of the criteria based content analysis and reality monitoring. *Applied Cognitive Psychology*, *33*(6), 1018–1031. https://doi.org/10.1002/acp.3537

Bogaard, G., Meijer, E. H., Vrij, A., & Nahari, G. (2022). Detecting deception using comparable truth baselines. *Psychology, Crime & Law*, *1-17*, 1–17. https://doi.org/10.1080/1068316X.2022.2030334

Bond, C. F., & DePaulo, B. M. (2008). Individual differences in judging deception: Accuracy and bias. *Psychological Bulletin*, *134*(4), 477–492. https://doi.org/10.1037/0033-2909.134.4.477

Bond, C. F., Jr., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review*, *10*(3), 214–234. https://doi.org/10.1207/s15327957pspr1003_2

Brandt, D. R., Miller, G. R., & Hocking, J. E. (1980). The truth-deception attribution: Effects of familiarity on the ability of observers to detect deception. *Human Communication Research*, *6*(2), 99–110. https://doi.org/10.1111/j.1468-2958.1980.tb00130.x

Brandt, D. R., Miller, G. R., & Hocking, J. E. (1982). Familiarity and lie detection: A replication and extension. *Western Journal of Communication (includes communication reports)*, *46*(3), 276–290. https://doi.org/10.1080/10570318209374086

Caso, L., Palena, N., Carlessi, E., & Vrij, A. (2019). Police accuracy in truth/lie detection when judging baseline interviews. *Psychiatry, Psychology and Law*, *26*(6), 841–850. https://doi.org/10.1080/13218719.2019.1642258

Caso, L., Palena, N., Vrij, A., & Gnisci, A. (2019). Observers' performance at evaluating truthfulness when provided with comparable truth or small talk baselines. *Psychiatry, Psychology and Law*, *26*(4), 571–579. https://doi.org/10.1080/13218719.2018.1553471

Colwell, K., Hiscock, C. K., & Memon, A. (2002). Interviewing techniques and the assessment of statement credibility. *Applied Cognitive Psychology*, *16*(3), 287–300. https://doi.org/10.1002/acp.788

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., &Cooper, H. (2003). Cues to deception. *Psychological Bulletin*, *129*(1), 74–118. https://doi.org/10.1037/0033-2909.129.1.74

Ewens, S., Vrij, A., Jang, M., & Jo, E. (2014). Drop the small talk when establishing baseline behaviour in interviews. *Journal of Investigative Psychology and Offender Profiling*, *11*(3), 244–252. https://doi.org/10.1002/jip.1414

Hartwig, M., & Bond, C. F., Jr. (2011). Why do lie-catchers fail? A lens model meta-analysis of human lie judgments. *Psychological Bulletin*, *137*(4), 643–659. https://doi.org/10.1037/a0023589

Hartwig, M., & Bond, C. F. (2014). Lie detection from multiple cues: A meta-analysis. *Applied Cognitive Psychology*, *28*(5), 661–676. https://doi.org/10.1002/acp.3052

Hauch, V., Sporer, S. L., Masip, J., & Blandón-Gitlin, I. (2017). Can credibility criteria be assessed reliably? A meta-analysis of criteria-based content analysis. *Psychol Assess*, *29*(6), 819–834. https://doi.org/10.1037/pas0000426

Ioannou, M., & Hammond, L. (2015). The detection of deception within investigative contexts: Key challenges and core issues. *Journal of Investigative Psychology and Offender Profiling*, *12*(2), 107–118. https://doi.org/10.1002/jip.1433

Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes factors. *The Journal of Problem Solving*, *7*, 2–9. https://doi.org/10.7771/1932-6246.1167

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. *Psychological Review*, *88*, 67–85.

Kashy, D. A., & DePaulo, B. M. (1996). Who lies? *Journal of Personality and Social Psychology*, *70*(5), 1037–1051. https://doi.org/10.1037/0022-3514.70.5.1037

Lee, M. D., & Wagenmakers, E. J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Levine, T. R. (2014). Truth-default theory (TDT) a theory of human deception and deception detection. *Journal of Language and Social Psychology*, *33*(4), 378–392. https://doi.org/10.1177/0261927X14535916

Levine, T. R., & Daiku, Y. (2019). How custom agents really detect lies. *Communication Research Reports*, *36*(1), 84–92. https://doi.org/10.1080/08824096.2018.1555523

Luke, T. J. (2019). Lessons from Pinocchio: Cues to deception may be highly exaggerated. *Perspectives in Psychological Science*, *14*(4), 646–671. https://doi.org/10.1177/1745691619838258

Mann, S., Vrij, A., & Bull, R. (2004). Detecting true lies: Police officers' ability to detect suspects' lies. *Journal of Applied Psychology*, *89*(1), 137–149. https://doi.org/10.1037/0021-9010.89.1.137

Masip, J., Sporer, S. L., Garrido, E., & Herrero, C. (2005). The detection of deception with the reality monitoring approach: A review of the empirical evidence. *Psychology Crime & Law*, *11*(1), 99–122. https://doi.org/10.1080/10683160410001726356

Merckelbach, H. (2004). Telling a good story: Fantasy proneness and the quality of fabricated memories. *Personality and Individual Differences*, *37*(7), 1371–1382. https://doi.org/10.1016/j.paid.2004.01.007

Moston, S., & Engelberg, T. (1993). Police questioning techniques in tape recorded interviews with criminal suspects. *Policing and Society: An International Journal*, *3*(3), 223–237. https://doi.org/10.1080/10439463.1993.9964670

Nahari, G. (2019). Verifiability approach: Applications in different judgmental settings. In *The Palgrave handbook of deceptive communication* (pp. 213–225). Springer.

Nahari, G., & Pazuelo, M. (2015). Telling a convincing story: Richness in detail as a function of gender and information. *Journal of Applied Research in Memory and Cognition*, *4*(4), 363–367. https://doi.org/10.1016/j.jarmac.2015.08.005

Nahari, G., & Vrij, A. (2015). Systematic errors (biases) in applying verbal lie detection tools: Richness in detail as a test case. *Crime Psychology Review*, *1*, 98–107. https://doi.org/10.1080/23744006.2016.1158509

Oberlader, V. A., Naefgen, C., Koppehele-Gossel, J., Quinten, L., Banse, R., & Schmidt, A. F. (2016). Validity of content-based techniques to distinguish true and fabricated statements: A meta-analysis.

Law and Human Behavior, *40*(4), 440–457. https://doi.org/10.1037/lhb0000193

Palena, N., Caso, L., Vrij, A., & Orthey, R. (2018). Detecting deception through small talk and comparable truth baselines. *Journal of Investigative Psychology and Offender Profiling*, *15*(2), 124–132. https://doi.org/10.1002/jip.1495

Schelleman-Offermans, K., & Merckelbach, H. (2010). Fantasy proneness as a confounder of verbal lie detection tools. *Journal of Investigative Psychology and Offender Profiling*, *7*(3), 247–260. https://doi.org/10.1002/jip.121

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137–149. https://doi.org/10.3758/BF03207704

Steller, M., & Köhnken, G. (1989). Criteria based statement analysis. In D. C. Raskin (Ed.), *Psychological methods in criminal investigation and evidence* (pp. 217–245). Springer.

Strömwall, L. A., & Granhag, P. A. (2003). How to detect deception? Arresting the beliefs of police officers, prosecutors and judges. *Psychology, Crime and Law*, *9*, 19–36. https://doi.org/10.1080/10683160308138

Tomas, F., Dodier, O., & Demarchi, S. (2021). Baselining affects the production of deceptive narratives. *Applied Cognitive Psychology*, *35*(1), 300–307. https://doi.org/10.1002/acp.3768

Verigin, B. L., Meijer, E. H., & Vrij, A. (2020). A within-statement baseline comparison for detecting lies. *Psychiatry, Psychology and Law*, *1-10*, 94–103. https://doi.org/10.1080/13218719.2020.1767712

Verigin, B. L., Meijer, E. H., Vrij, A., & Zauzig, L. (2019). The interaction of truthful and deceptive information. *Psychology, Crime & Law*, *367-383*, 367–383. https://doi.org/10.1080/1068316X.2019.1669596

Vrij, A. (2008b). Nonverbal dominance versus verbal accuracy in lie detection—A plea to change police practice. *Criminal Justice and Behavior*, *35*(10), 1323–1336. https://doi.org/10.1177/0093854808321530

Vrij, A. (2014). Interviewing to detect deception. *European Psychologist*, *19*, 184–194.

Vrij, A. (2016). Baselining as a lie detection method. *Applied Cognitive Psychology*, *30*(6), 1112–1119. https://doi.org/10.1002/acp.3288

Vrij, A., Akehurst, L., Soukara, S., & Bull, R. (2002). Will the truth come out? The effect of deception, age, status, coaching, and social skills on CBCA scores. *Law and Human Behavior*, *26*(3), 261–283. https://doi.org/10.1023/a:1015313120905

Vrij, A., Akehurst, L., Soukara, S., & Bull, R. (2004). Let me inform you how to tell a convincing story: CBCA and reality monitoring scores as a function of age, coaching and deception. *Canadian Journal of Behavioural Science*, *36*, 113–126. https://doi.org/10.1037/h0087222

Vrij, A., Fisher, R. P., & Blank, H. (2017). A cognitive approach to lie detection: A meta-analysis. *Legal and Criminological Psychology*, *22*(1), 1–21. https://doi.org/10.1111/lcrp.12088

Wagenmakers, E. J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Selker, R., Gronau, Q. F., Dropmann, D., Boutin, B., Meerhoff, F., Knight, P., Raj, A., van Kesteren, E. J., van Doorn, J., Smira, M., Epskamp, S., Etz, A., Matzke, D., … Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*, *25*(1), 58–76. https://doi.org/10.3758/s13423-017-1323-7

How to cite this article: Bogaard, G., Nußbaum, M., Schlaudt, L. S., Meijer, E. H., Nahari, G., & Vrij, A. (2022). A comparable truth baseline improves truth/lie discrimination. *Applied Cognitive Psychology*, *36*(5), 1060–1071. https://doi.org/10.1002/acp.3990

APPENDIX

Decision Rule

To test our decision rule 'if the baseline is more detailed than the target, it is likely that the interviewee is lying and if the baseline is less detailed than the target, it is likely that the interviewee is telling the truth', we created a new variable 'Decision Rule'. When the RM difference score ≤0, Decision Rule was assigned the value 1 (indicating truth), when the RM difference score was >0, Decision Rule was assigned the value 2 (indicating deception). Next, we conducted a Chi-square test with Veracity and Decision Rule. Results showed there was a significant association between the two variables, $X^2(1, 87) = 5.97$, $p = .02$. The proportion of lies that were correctly classified using this rule was 25/41 (61%), and the proportion of truths was 30/46 (65%).