UvA-DARE (Digital Academic Repository)

The verifiability approach to deception detection: A preregistered direct replication of the information protocol condition of Nahari, Vrij, and Fisher (2014b)

Verschuere, B.; Schutte, M.; van Opzeeland, S.; Kool, I.

Link to publication

Citation for published version (APA):

SHORT PAPER

WILEY

The verifiability approach to deception detection: A preregistered direct replication of the information protocol condition of Nahari, Vrij, and Fisher (2014b)

Bruno Verschuere ⬤ | Manon Schutte | Sharon van Opzeeland | Ilona Kool

Department of Clinical Psychology, University of Amsterdam, Amsterdam, Netherlands

Correspondence
Bruno Verschuere, Department of Clinical Psychology, University of Amsterdam, Amsterdam, Netherlands.
Email: b.j.verschuere@uva.nl

Summary

Nahari, Vrij, and Fischer [(2014b), *Applied Cognitive Psychology*, *28*, 122–128] found that, when participants were forewarned that their statements would be checked for verifiable details, truth tellers gave much more verifiable details than liars. In this direct replication (*n* = 72), participants wrote a statement claiming they had carried out their regular campus activities, whereas liars had actually stolen an exam. Statements were coded for verifiable details. Our primary prediction was confirmed: Truth tellers provided significantly more verifiable details than liars. Of note, the replication effect size (*d* = 0.49) was less than half that of the original (*d* = 1.14), and – like in the original study – was smaller than the lie-truth effect size for total details (verifiable and unverifiable details combined; *d* = 0.80). We hope this will stimulate other independent investigations of VA to tell whether or not coding for verifiability will pass Ockham's razor test.

KEYWORDS

deception detection, direct replication, reality monitoring, replication, verifiability approach

1 | INTRODUCTION

The Verifiability Approach (VA; Nahari et al., 2014a, 2014b; for a recent review see Vrij & Nahari, 2019) is a promising new lie detection approach. It builds on the idea that truth tellers are typically willing and able to provide a narrative rich in detail whereas liars are in a dilemma when telling their story. On the one hand, liars are reluctant to talk, as anything they say could be used to uncover their deceit. On the other hand, liars realize they have to provide a detailed account to come across as truthful. The solution to this dilemma may be to provide details that cannot be verified, and to refrain from providing details that can be checked (e.g., "In the park, I just lit a Marlboro Light. Startled by a rabbit passing by, I dropped my fag on the wet grass. I swore because it was the last cigarette of

my package! Can you imagine my frustration?"). To discern lie from truth, the VA therefore uses a two-step procedure in coding statements. Building on Reality Monitoring theory (Johnson & Raye, 1981) and a substantial number of empirical studies (Masip et al., 2005), the statement is first coded for the presence of perceptual (i.e., things the person saw, enacted, heard, tasted, or smelled) and contextual details (i.e., information about locations or the spatial arrangement of people and/or objects; information about when the event happened, the duration of an activity, or information about the sequence of events). Then, these details are coded as being verifiable or unverifiable. A detail is considered verifiable if its truthfulness can potentially be checked because the activity was carried out with or witnessed by an identifiable other, or because it was documented (e.g., through using a credit card or smartphone) or recorded (e.g., by security camera). The main prediction of the VA is that truth tellers are predicted to provide more verifiable details than liars.

Bruno Verschuere and Manon Schutte shared first authorship.

The VA has received substantial attention, from both scholars and practitioners. The 2014a paper has been cited more than 100 times on Google Scholar in just a few years. VA prominently features in leading reviews on investigative interviewing (e.g., Vrij et al., 2017). That the UK Centre for Research and Evidence on Security Threats (CREST) issued a VA guide for interviewers illustrates the applied interest (see https://crestresearch.ac.uk/resources/liars-struggle-to-provide-checkable-details/). The VA is indeed appealing in several ways. With a single cue (i.e., the number of verifiable details), large effects have been reported (e.g., Harvey et al., 2017; Nahari et al., 2014b). Moreover, while lie detection tools are notably vulnerable to faking (Honts, 2014), VA would not just be resistant to but in fact benefit from informing the examinee what the tool is about. Indeed, giving a forewarning would stimulate truth tellers, but not liars, to provide more verifiable details (Nahari et al., 2014b). The VA also provides for a much-needed within-person index. While most verbal lie detection tools do not provide guidance for decision making in individual cases, VA has proposed the ratio of verifiable details to non-verifiable details (or to total number of details) as a within-person index (Nahari & Vrij, 2019).

At the same time, there is ample reason to be cautious. First, there is a lack of independent VA research. A recent review found 13 published reports on the VA (Bogaard et al., 2020). All (100%) of those studies had at least one of the three original authors (Nahari, Vrij, Fisher) as co-author. Studies co-authored by tool designers can lead to larger effects than those by independent evaluations (Singh et al., 2013). Second, the results of empirical studies on VA are mixed, with some studies finding no effect (e.g., Boskovic et al., 2019; $d = -0.03$), others small effects (e.g., Jupe et al., 2017; $d = 0.28$) and still others finding large effects (e.g., Model Statement condition of Harvey et al. (2017); $d = 1.22$). Several moderators have been proposed, including the scenario (with most promising results for criminal and insurance settings) and the use of forewarning (i.e., called the "information protocol") (Nahari & Vrij, 2019). The use of forewarning is now recommended by the original VA authors, and, for some settings, even considered crucial. Third, large effects for single cues are thus far only found in initial studies and a negative correlation between the amount of studies and the effect size has been found (Bond et al., 2015).

Because there exists considerable uncertainty on the lie-truth effect sizes obtained with the VA, we decided to conduct a pre-registered, direct, independent replication of the information protocol condition reported in Nahari et al. (2014b). In this mock crime study, liars stole an exam on a university campus, whereas truth tellers were free to do their regular business on the university campus. Both liars and truth tellers provided a statement claiming they had been doing their normal business on campus. Particularly after a forewarning that VA would be applied, truth tellers were found to include much more verifiable details in their statement than liars ($d = 1.14$). We chose this design because effects are theorized to vary with scenario. The mock crime scenario has been argued to be particularly suited for VA, as truth tellers will be able to provide more verifiable details of the alibi than liars (Bogaard et al., 2020; Vrij & Nahari, 2019). Given the applied purpose of VA, we chose for the condition that was reasoned and found to produce the maximal lie-truth effect (i.e., the use of forewarning). The primary target effect for our replication

effort was that truth tellers would provide more verifiable details than liars, and we defined replication success as a statistically significant effect in the same direction as the original study. We also pre-registered the secondary hypothesis that liars will provide more non-verifiable details than truth-tellers.[1] Finally, we pre-registered the secondary hypothesis that the ratio of verifiable details to non-verifiable details would be higher for truth tellers than for liars.

2 | REPRODUCING THE ORIGINAL EFFECT OF INTERESTS

Based on the data provided by the first author of the original paper, we were able to reproduce the primary effect of interest: Truth tellers ($M = 17.86$, $SD = 11.54$) provided more verifiable details than liars ($M = 7.04$, $SD = 6.90$). This was a significant and large effect $d = 1.14$ (95% CI: 0.49; 1.78). A Bayesian independent sample t-test, using a zero-centered Cauchy prior (r) scaled at 0.707 (the default setting in JASP 0.12.2.0; JASP Team, 2019), showed that the data were 51.35 times more likely under the hypothesis that there is a difference in amount of verifiable details between liars than truth tellers than under the null hypothesis of no difference.

3 | METHOD

3.1 | Pre-registration

The study was pre-registered at Open Science Framework before data collection (https://osf.io/ucz92). The study was approved by the Ethics Review Board of the University of Amsterdam (archived as 2019-CP-11275).

3.2 | Deviations from Nahari et al., 2014b

We aimed to directly replicate the method of Nahari et al., 2014b. Our replication design deviated from the original on the following aspects: We ran the study in the Netherlands (original: Israel), and awarded an additional possible bonus for providing a credible statement of 35 euro (original: 150 Israeli Shekels). As in the original study, we tested participants in their native language (original: Hebrew; here: Dutch). As in the original study, we matched the time for truth tellers to the time that liars needed to enact the mock theft. In the original study, liars required 30 min. Due to differences in campus size and organization, liars required less time in the replication study to complete the same mock crime. In agreement with the first author of the original paper, we therefore changed the time from 30 to 15 min for truth tellers to engage in their regular campus activities (but see Discussion). Finally, we added the exclusion criterium "not following instructions," and we added one question to the post-statement questionnaire ("What do you think the odds are the interviewer will verify the details in your statement?").

3.3 | Deviations from pre-registration

The sampling plan in the pre-registration stated that data collection would stop after 70 inclusions. Due to additional planned appointments, we ended up with 72 inclusions.

3.4 | Participants

The effect of interest was of size $d = 1.14$ in the original study. Only 36 participants would be needed to pick up such a directional effect with 95% power. Original studies often overestimate the true effect size, with replication studies on average finding a 50% reduction for successful replications (Camerer et al., 2018). A realistic power analysis can therefore better take into account a smaller than original true effect size. As lie detection tools require large effects, we chose to be able to detect an effect of at least $d = 0.80$ (acknowledging that smaller effects could be of theoretical interest while larger effects would be preferred for VA to be useful as a stand-alone lie detection tool; see also the Limitations section of the Discussion). A power analysis using GPower 3.1 (Erdfelder et al., 1996), with $\alpha = 0.05$, indicated that 70 participants were needed to be able to pick up such an effect in a one-tailed, independent sample t-test with 95% power.

Participants were eligible for inclusion if they were fluent in Dutch. Seventy-three participants (M age = 22.34, SD = 7.79; 53 females) engaged in the experiment, in exchange for money (€7.50) or participation credits. Exclusion criteria consisted of not following the instructions correctly or not providing a statement. This was the case for one participant, who, for technical reasons, could not complete the mock crime. Thus, the final sample consisted of 72 participants (M age = 22.39, SD = 7.84; 52 females): 38 truth-tellers and 34 liars.

3.5 | Procedure

A detailed description of the procedure, including the instructions (the English translation of the original study instructions obtained from Nahari et al., 2014b, as well as our Dutch adaptation) can be found on https://osf.io/3zk59/files/. We followed the original instructions (including those for the Information Protocol) as close as possible, making only the necessary adjustments for our mock crime.

After reading and signing the informed consent, the first experimenter instructed the participant according to their random assignment to either the lie condition or the truth tellers condition. The liars were told to carry out a mock crime, which consisted of stealing and copying a statistics exam from the department's mail box. The lab, the exam, and the copier were in different buildings. The truth tellers were instructed to carry out their normal activities around campus at different locations, to spend a maximum of 5 min in each location, for a total duration of 15 min. The instructions were repeated and the experimenter made sure any remaining questions were addressed. Once returned to the lab, participants were told by the experimenter that during their absence an exam had been stolen and they were suspected of this theft. In order to prove their innocence, the participant had to provide a statement about their whereabouts in the last 15 min. Guilty participants were told to not confess. All participants were informed that the interviewer who would read their statements would check whether or not the details in the statement could be verified, with an explanation of verifiable details. Participants were instructed to be as convincing as possible, and were told that if the interviewer believed they were innocent, they would participate in a draw in which four participants could earn an extra €35 each. If the interviewer would not believe them, the participant would not partake in the draw and would have to write an additional statement.[2] After repeating the instructions and assuring everything was clear, the first experimenter gave the participant 10 min to prepare their statement. After these 10 min, the participant filled in a pre-statement questionnaire about their planned strategy (https://osf.io/56mkn/).

The first experimenter then brought the participant to a different room, and left the participant there with the second experimenter (the interviewer who was blind to veracity). The second experimenter asked the participant to provide a handwritten statement of their activities in the last 15 min. Specifically, the second experimenter informed the participants that "An exam was stolen from the mailbox in the G-building in the last hour. You are one of the suspects of this theft. Please describe in as much detail as possible what you have been doing on campus in the last 15 minutes so I can make sure you have an alibi and that you didn't steal the exam. Start your statement from the moment you left the lab and end when you got back to the lab. This is the only question I will ask you. You have this one chance to convince me that you didn't steal the exam."

After completing their statement, the participant was brought back to the experimenter room and was asked to fill in a post-statement questionnaire (https://osf.io/n7zmg/). Finally, the participant received a copy of their own statement and was given a marker to highlight all the correct details in their statement (veracity manipulation check).

3.6 | Coding

We (BV and MS) developed the coding scheme based on the VA literature, a VA coding workshop by the first author of the original paper, and extensive discussions with the first author of the original paper. We piloted this study (one truthful, and one deceptive participant) and the VA coding of these pilot statements, then fine-tuned and finalized the coding scheme. We pre-registered the VA coding scheme on https://osf.io/whkcz/ before data collection.

Each statement was coded, independently and blind to veracity condition by two coders (IK and SVO). The coders were acquainted with the VA literature and received the detailed VA coding scheme. In a 2 h training session (led by BV and MS), questions on the coding scheme were discussed, and two example statements, obtained from pilot participants of this study, were coded independently and blind to veracity. The coding of all statements was discussed with and disagreements were solved by a third coder (MS). The two main coders achieved high rates of agreeability, with an ICC of 0.93 for verifiable details, 0.89 for unverifiable details, and 0.87 for total number of details

(all p < .001). These values were calculated with the ICC function in the psych R package (W Revelle, 2015; William Revelle, 2012) and we report the value for ICC3 (a fixed set of k judges rate each target using the formula ((MSB - MSE)/(MSB+ (nr−1)*MSE), with MSB, MSE and nr referring to Mean Square Between subjects, Mean Squared Error (within subjects), and number of decisions, respectively).

The detailed coding scheme is available on https://osf.io/whkcz/. In brief, statements were coded on spatial details (i.e., information about spatial arrangements of things, people and locations), perceptual details (i.e., details about what the participant saw, felt, heard or smelt) and temporal details (i.e., about the order of events or the timing of an event), and each of these details were then coded as a verifiable detail or an unverifiable detail. Verifiable details were defined as "perceptual and contextual details" (i.e., the spatial and temporal details) related to: (1) activities that were carried out with named persons or persons who can be identified based on the description given (2) activities that were witnessed by named persons or identifiable persons, or (3) activities that were documented, recorded or that the interviewee believes may have been captured on CCTV. Details that did not meet these conditions were coded as being unverifiable.

We illustrate our VA coding through a section of a statement by a pilot participant claiming to have performed his regular campus activities. On the left is the statement. On the right is the coding of the statement. Note that not all parts of the statement are coded, and *in italic* we explain why an information unit is not counted as a detail. The parts **in bold** were counted as a (contextual or perceptual) detail. When underlined, the detail was subsequently coded as being verifiable (The person at the counter has witnessed these actions. Because we know the place and time of the encounter, that person is identifiable).

Because I have not eaten anything for breakfast yet	*Reasoning*
I looked through the food that is offered at the cafeteria.	1 Perceptual detail
	1 Spatial detail
I never went there so I was a bit overwhelmed at first.	*Out of context + Emotion*
After checking whether or not they have vegan offers	1 Temporal detail
	1 Perceptual detail
(which they do at some of the stands)	*Out of context*
I chose a tempeh	1 Perceptual detail
sandwich	1 Perceptual detail
at tashas	1 Spatial detail
(not sure about the name anymore since it was my first time eating there)	*Out of context*
The person at the counter	1 Perceptual detail
said it is going to take 5 minutes to prepare the sandwich.	1 Perceptual detail

3.7 | Materials

Participants filled in a short questionnaire before and after giving their statement. These are the same questionnaires used in the original study, translated to Dutch. The pre-statement questionnaire (original: https://osf.io/zuec9/; Dutch translation: https://osf.io/56mkn/) asked 3 questions about level of preparation and 1 on level of motivation, all on 7-point Likert scale. Higher scores indicate better preparation/motivation, see Table 2. Participants were also asked whether or not they prepared a strategy to appear credible. We did not analyze the two open questions. The post-statement questionnaire (original: https://osf.io/rwdby/; Dutch translation: https://osf.io/n7zmg/) asked 6 questions, all on 7-point Likert scale (from not at all [likely] to very much/likely), about motivation, difficulty, credibility, likelihood reward and negative consequences, and our additional question about the odds that details would indeed be verified, see Table 3.

4 | RESULTS

4.1 | Confirmatory analyses

We used three one-tailed independent sample Welch t-tests for the confirmatory analyses. The primary hypothesis was confirmed: Truth tellers provided significantly more verifiable details than liars, t(59.22) = 2.17, p = .017, d = 0.50 (95%CI: 0.02, 0.98), see Table 1. The secondary hypothesis that liars would tell more unverifiable details than truth tellers was not confirmed, and numerically in the opposite direction, t(67.10) = −1.65, p = .947, d = −0.38 (95% CI: −0.86, 0.09). The secondary hypothesis that the ratio verifiable to unverifiable details would be higher for truth tellers than for liars was also not confirmed, t(37.70) = −1.09, p = .142, d = 0.24 (95% CI: −0.23, 0.72).[3] The data and analyses scripts can be found on https://osf.io/3zk59/.

4.2 | Exploratory analyses

4.2.1 | Benchmarks for interpreting the VA effect size

A one-tailed independent sample Welch t-test showed that the total amount of details (i.e., verifiable and unverifiable details combined) in truthful statements was significantly higher compared to deceptive statements, t(68.58) = 3.43, p < .001, d = 0.80 (95% CI: 0.31, 1.29), see Table 1. A one-tailed independent sample Welch t-test showed that statements of truth tellers were not significantly longer than statements of liars, t(69.59) = 1.55, p = .063, d = 0.36 (95%CI: −0.11, 0.84), see Table 1.

4.2.2 | Bayesian statistics

A one-tailed Bayesian independent sample t-test, using the JASP default settings, showed that the data were about 3 times more likely

TABLE 1 Means (SDs) for the measured variables for truth tellers and liars

Status of hypothesis	Dependent variable	Truth tellers M (SD)	Liars M (SD)	Effect size Cohen's d [95% CI]	Bayes factor (BF$_{+0}$)
Primary hypothesis	Verifiable details	7.18 (8.35)	3.76 (4.67)	0.50[4] [0.02, 0.98]	3.11
Secondary hypotheses	Unverifiable details	17.63 (11.48)	13.79 (8.28)	- 0.38 [−0.86, 0.09]	0.10
	Ratio verifiable to unverifiable details[5]	1.72 (6.92)	0.50 (0.64)	0.24 [−0.23, 0.72]	0.64
Other (benchmark) effects of interest	Total number of details	24.82 (10.11)	17.56 (7.80)	0.80 [0.31, 1.29]	52.83
	Number of words	183.24 (95.39)	151.38 (78.92)	0.36 [−0.11, 0.84]	1.22

Note: BF$_{+0}$ denotes how much more likely the data are under the hypothesis of lie-truth differences (one-sided) compared to the null hypothesis of no lie-truth differences.

TABLE 2 Means (SDs) for the pre-statement questions for truth tellers and liars

Question	Truth tellers M (SD)	Liars M (SD)	Effect size cohen's d [95% CI]	Bayes factor (BF$_{10}$)
How would you describe the preparation for your statement (1 = superficial to 7 = thorough)?	4.79 (1.44)	4.68 (0.88)	0.09 [−0.37, 0.56]	0.26
How would you describe the preparation for your statement (1 = insufficient to 7 = sufficient)?	5.71 (0.90)	5.09 (1.24)	0.58 [0.12, 1.05]	3.07
How would you describe the preparation for your statement (1 = bad to 7 = good)	5.26 (0.92)	4.82 (1.09)	0.44 [−0.03, 0.91]	1.05
How motivated are you to come across as credible (1 = not at all to 7 = very much)	5.45 (1.03)	5.68 (0.95)	0.23 [−0.23, 0.69]	0.41
Do you have a strategy to come across as credible - % YES	73.68%	82.35%	0.10 (Cramer's V)	0.35

TABLE 3 Means (SDs) for the post-statement questions for truth tellers and liars

Question (all scored from 1 = not at all to 7 = very much)	Truth tellers M (SD)	Liars M (SD)	Effect size cohen's d [95% CI]	Bayes factor (BF$_{10}$)
How motivated were you to come across as credible?	5.97 (0.89)	5.97 (0.76)	0.004 [−0.46, 0.47]	0.24
How difficult was it to come across as credible?	3.37 (1.60)	4.76 (1.44)	0.92 [0.425, 1.40]	105.26
How credible do you think you were?	5.34 (1.15)	4.50 (0.96)	0.79 [0.31, 1.27]	24.87
How likely do you think the chance is that you will participate in the draw?	4.79 (1.44)	3.76 (1.33)	0.74 [0.26, 1.22]	14.04
How likely do you think the chance is that you will have to write a second statement?	3.16 (1.44)	4.44 (1.11)	0.99 [0.50, 1.48]	275.95
How likely do you think the chance is that the interviewer will check the details in your statement?	4.18 (1.71)	3.85 (1.94)	0.18 [−0.28, 0.65]	0.31

under the hypothesis that truthful statements contained more verifiable details than deceptive statements than under the null hypothesis of no difference, see Table 1. This fits with the wide confidence interval around the obtained effect, and indicates that the evidential strength of the data was only modest. Table 1 further shows the Bayes Factors that we obtained for directional tests for the other dependent variables. For unverifiable details, the data were 10 times more likely under the null hypothesis than under the hypothesis that liars would provide more unverifiable details than truth tellers. For the ratio score and word count, the data were inconclusive and were equally likely under the alternative hypothesis of a lie-truth difference as under the null hypothesis of no lie-truth difference. For total details

(verifiable and unverifiable details combined), the data were 52.83 times more likely under the hypothesis of truthful statements being more detailed compared to deceptive statements than under the null hypothesis of no difference. In other words, the evidential strength for total details (BF_{+0} = 52.83) was substantially stronger than that for verifiable details (BF_{+0} = 3.11).

4.3 | Impact of outliers

For verifiable details, there was one outlier. One (truthful) statement contained 43 verifiable details, thereby deviating more than 6 SDs from the mean. Excluding this outlier, the effect for verifiable details was, $t(67.46)$ = 1.95, p = .028, d = 0.46 (95%CI: −0.02, 0.93), BF_{+0} = 2.26. For unverifiable details, there was one outlier. One (truthful) statement contained 65 unverifiable details, thereby deviating more than 3 SDs from the mean. Excluding this outlier, the effect for unverifiable details was, $t(68.71)$ = −1.29, p = .899, d = −0.31 (95% CI: −0.16, 0.77), BF_{+0} = 0.12. For the ratio score, there was one outlier. One (truthful) statement had a ratio score of 43, thereby deviating more than 8 SDs from the mean. Excluding this outlier, the effect for the ratio score was, $t(67.82)$ = −0.65, p = .515, d = 0.15 (95% CI: −0.31, 0.62), BF_{+0} = 0.43. In sum, excluding outliers did not impact the conclusions with regard to our primary (p < .05) nor our secondary hypotheses (p's > .05). The data do further strengthen the impression that the evidential strength for the primary prediction is modest.

4.4 | Reformulation of the secondary hypothesis regarding unverifiable details

The VA predicts liars to provide unverifiable details (Nahari et al., 2014a; Vrij & Nahari, 2019). Like others (Bogaard et al., 2020; Harvey et al., 2017), we predicted that liars would provide more unverifiable details than truth tellers. Liars may not necessarily make more unverifiable details compared to truth tellers, but rather compared to verifiable details (Nahari, personal communication, May 14, 2020). We tested this reformulated prediction using a paired sample t-test. Liars indeed told more unverifiable than verifiable details, d = 0.92 (95%CI: 0.51; 1.31), BF_{10} = 3005. But importantly, the same is true for truth tellers, d = 0.60 (95%CI: 0.25; 0.94), BF_{10} = 44.18. These effects likely result from the stricter conditions to code a detail as verifiable than as unverifiable, hampering the falsifiability of the reformulated hypothesis.

Vrij and Nahari (2019) note that it is more difficult to make predictions regarding the number of unverifiable details, and like Harvey et al. (2017), they speculated that the effect for unverifiable details may depend on the motivation of the liar to provide a detailed statement. This is an interesting possible moderator, that does require a criterion for motivation that is independent of the outcome. Lacking such a criterion, there is a risk for circularity, that is, that liars were

(not) motivated if they (do not) provide more unverifiable details. In the current study, we found no relation between motivation and number of unverifiable details in liars (motivation assessed pre-statement: r = −0.18, p = .30; motivation assessed post-statement: r = 0.10, p = 0.57).

4.5 | Veracity manipulation check

Following Nahari et al. (2014b), we checked condition assignment by examining the proportion of the statement identified by the participant as truthful. The data of two participants were lost for this check. Truth tellers marked on average 67.32% (SD = 34.42) of their statement as truthful and liars marked on average 31.91% (SD = 27.68) of their statement as truthful, and this effect was significant, independent sample Welch $t(68)$ = 4.77, p < .001, d = 1.12 (95% CI: 0.62, 1.64), BF_{+0} = 2512.16. Of note is that 7 out of 32 liars (22%) marked more than 50% of the details in their statement as truthful, but particularly that 13 out of 38 truth tellers (34%) marked less than 50% of the details as truthful. Particularly the latter finding might be due to the ambiguous task instructions - prompting participants to mark *details* that were truthful, rather than simply asking to mark everything that was truthful. While we deem it very unlikely, we cannot exclude that some truthful participants were partly deceptive. In a sensitivity analysis, we therefore included only participants who marked more than half of their provided details in line with their assigned condition. This left our study conclusions unchanged: The lie-truth effect for verifiable details was d = .62 (95% CI: 0.03; 1.20), for unverifiable details, d = −0.48 (95% CI: −1.06; 1.09), for the ratio score d = 0.31 (95% CI: −0.26; 0.88), for total number of details d = 1.03 (95% CI: 0.43; 1.64), and for word count d = 0.68 (95% CI: 0.09; 1.26).

4.6 | Pre-statement questionnaire

Inspection of Table 2 shows that participants reported to be well prepared for their statement, to be motivated to appear credible, and that most participants had a strategy. For sake of completion, we also report differences between liars and truth tellers (Cohen's d and BF_{10}). There were no meaningful differences between liars and truth tellers.

4.7 | Post-statement questionnaire

Table 3 reports the results for the questionnaire completed after providing the statement. Participants again indicated to have been motivated to appear credible. But, there were now also substantial differences between liars and truth tellers. Liars found it more difficult to appear credible and thought they were less credible, compared to truth tellers. Compared to truth tellers, liars also thought it less likely they would participate in the draw (for appearing credible) and more likely they would have to write a second statement (for not appearing

credible). Of note is also that only 46% of participants (n = 33; 18 truth tellers and 15 liars) thought it was likely that their details would actually be checked (i.e., they gave a score of 5 or higher). When restricting the analyses to this subsample, the lie-truth effect for verifiable details was d = 0.64 (95% CI: −0.09; 1.38), for unverifiable details, d = 0.49 (95% CI: −0.23; 1.22), for the ratio verifiable details d = 0.08 (95% CI: −0.64; 0.79), for total number of details d = 0.89 (95% CI: 0.15; 1.64), and for word count d = 0.26 (95% CI: −0.46; 0.97).

5 | DISCUSSION

Lie detection is a challenging task. Most people are very poor in telling lie from truth, and most non-verbal and verbal cues to deception are faint (DePaulo et al., 2003; Luke, 2018). Nahari et al. (2014b) found initial support showing there may be promise to a novel, single cue - the amount of verifiable details. The results of later studies provided more diverse results, creating considerable uncertainty on the size of the effect. We set up a direct pre-registered replication to examine whether verifiable details provide for a reliable cue to deception.

Following our predefined replication criterion, we consider the replication successful: Being forewarned, truth tellers provided significantly more verifiable details than liars. This is encouraging because it is the main prediction of the VA (Vrij & Nahari, 2019), and this may be the first study that is fully independent of the original authors. Thus, there is promise to verifiable details as a cue to deceit. Our choice to rely on statistical significance as a criterion for success could, however, be questioned. Indeed, the imaginary case of an original study showing p = .049 and the replication study showing p = .51 (or vice versa), illustrates its hazardous to solely rely on statistical significance. Therefore, it is worthwhile to also look at other criteria (Open Science Collaboration, 2015). One such a criterion is a qualitative estimate of the replication team. Four researchers (BV, MS, IK, SVO) were involved in this replication, and all four considered the replication successful. Another possible criterion is to compare the effect sizes of the original and the replication study. Here, the results shed a different picture: The effect size of the replication study (d = 0.50 [95% CI: 0.02; 0.98]) is less than half that of the original study (d = 1.14 [95% CI: 0.49; 1.78]), the effect size of the original study falls outside of the 95% CI of the replication study, and the lower bound of the wide confidence interval around the replication effect size touches zero. For the replication study, the Bayesian analyses also found the evidential strength for the lie-truth difference in verifiable details to be modest.

Differences between the original study and the replication study might explain the difference in effect size. For instance, it is possible that the shorter time frame restricted the number of verifiable details truth tellers could provide.[6] This calls for studies examining the impact of the time frame on VA validity – examining whether longer time frames not only lead to longer statements, but also larger lie-truth differences – and points to a possible boundary condition of VA (i.e., a minimum time frame). Also, although a suboptimal measure, the veracity manipulation check indicated that the objective lie-truth

differences in the replication study may have been smaller compared to the original study. Exploratory analyses did not substantiate the idea this would have had substantial impact on the obtained effect size. And it is possible that the validity of VA coding by the original VA authors is higher than that by our coders. Alternatively, it is possible that the verifiable detail effect may be lower than estimated by the original study. First, large effects for single cues are rare and typically only found for understudied cues (Bond et al., 2015). Second, the results of large-scale replication efforts consistently indicate that effect sizes in independent evaluation are substantially smaller - often by 50%, as we also observed here (Camerer et al., 2018). Third, the pre-registration of our methods, coding scheme, and analyses restricts researchers degrees of freedom. Fourth, this is the first study that is fully independent of the original research team. Fifth, the replication effect size is more in line with the meta-analytic result of Luke (2018), who found an effect size for detailedness of d = 0.30, based on 24 estimates with N = 883.

A key assumption of the VA is that liars can solve their dilemma - needing to provide details to appear credible and being reluctant to provide details that could uncover their deceit - by providing details that cannot be checked. Our prediction - and not necessarily still that of the original authors - that liars would provide more unverifiable details than truth tellers was not confirmed. Thus, it remains unclear whether and when deceptive and truthful statements differ in number of unverifiable details.

The most diagnostic cue in the current study was the total amount of details (verifiable and unverifiable details combined). Of note, obtaining the data of the original study, we learned this was also true for the original study (verifiable details: d = 1.14; verifiable and unverifiable details combined: d = 1.56). This finding questions the added value of differentiating between verifiable and unverifiable details. To clarify whether and when coding for verifiability can have added value to the more simple coding for detailedness, it would be helpful if future VA studies would systematically report on a few benchmark cues (as is common in neuropsychology, Lange, 2020). Apart from verifiable details, unverifiable details, and the ratio score, candidate benchmark cues could be the total number of details and statement length. Without evidence for the added value of scoring for verifiability, the principle of simplicity known as Ockham's razor (Jefferys & Berger, 1992) would lead one to chose for detailedness.

This is study is not without its limitations. First, while backed up with an a priori power calculation, the sample size of our study was modest. This may partly explain the wide confidence intervals and the uncertainty in the effect size estimates. Second, Nahari et al. (2014b, p. 127) reasoned that participants in the original study may have not believed that their details would be actually checked. We added a question that prompted participants beliefs and indeed found that less than half of the sample found it likely that their details would be checked. While this measure brings about important new information, its results may not come as a surprise (for similar findings see Boskovic et al., 2017). Study participants are likely to realize that experimenters will not and even cannot obtain access to security camera footage, unlock and inspect smartphones, or trace and interview

witnesses. We encourage VA researchers to assess whether participants believe that their details could and would be checked. This provides VA researchers with ethical, practical, and privacy challenges of how to make it believable to participants that their statement may be fact-checked. Field or semi-field like studies could be of help. Third, while having a clear focus, the scope of the present replication was limited (e.g., we used a specific scenario, and we did not examine whether the presumed beneficial effect of the information protocol replicates).

In sum, we found that, after a forewarning, truthful statements indeed contain significantly more verifiable details than deceptive statements. As commonly observed in replication studies, the effect size of the replication was about half of the original effect size. We hope this will stimulate other independent investigations of VA to tell whether or not coding for verifiability will pass Ockham's razor test.

ACKNOWLEDGMENT

We thank Galit Nahari for extensive support with this study, including data sharing, discussion and feedback on the VA coding scheme, and clarifications of the original procedure.

CONFLICT OF INTEREST

Bruno Verschuere is a long standing collaborator of the first and second author of the original study.

DATA AVAILABILITY STATEMENT

All data, materials, and analytic scripts are publically available on https://osf.io/3zk59/.

ORCID

Bruno Verschuere https://orcid.org/0000-0002-6161-4415

ENDNOTES

[1] This was not an explicit prediction of Nahari et al., 2014b. It was based upon our understanding of VA (e.g., "liars will provide more perceptual, spatial, and temporal details that cannot be verified"; Nahari et al., 2014a, p. 229), a matter we address in the Exploratory Analyses and the Discussion.

[2] In reality, all participants were told during the debriefing that the interviewer believed them and all participated in the draw.

[3] Because the assumption of normality was violated, we repeated the analyses using non-parametric tests. The results for the Mann–Whitney U tests yielded the same conclusions and can be found in full here: https://osf.io/mpxqa/

[4] A very helpful source to interpret Cohen's d is https://rpsychologist.com/d3/cohend/, which suggests that with $d = 0.5$, the verifiable detail scores of 80.3% of liars and truth tellers will overlap, and there is a 63.8% chance that a randomly picked truth teller will have a higher score than a randomly picked liar.

[5] After 2014, most VA papers used a different ratio calculation, that is, the number of verifiable details to the total number of details. The results for this ratio score were quite similar, $t(69.57) = 0.98$, $p = .165$, $d = 0.23$ (95% CI: −0.24, 0.70), $BF_{+0} = 0.60$.

[6] Our statements were not shorter (see summary stats in Table 1) compared to those of the original study (deceptive statements: $M = 100$,

$SD = 40$; truthful statements: $M = 191$, $SD = 112$; data obtained from the original authors). But the number of verifiable details was about half of that found in the original study.

REFERENCES

Bogaard, G., Meijer, E. H., & Van der Plas, I. (2020). A model statement does not enhance the verifiability approach. *Applied Cognitive Psychology*, *34*, 96–105. https://doi.org/10.1002/acp.3596.

Bond, C. F., Levine, T. R., & Hartwig, M. (2015). New findings in non-verbal lie detection. Granhag P. A., Vrij A, Verschuere B, In *Detecting deception: Current challenges and cognitive approaches*. Hoboken, NJ: John Wiley & Sons, Ltd. Retrieved from https://doi.org/10.1002/9781118510001.ch2.

Boskovic, I., Bogaard, G., Merckelbach, H., Vrij, A., & Hope, L. (2017). The Verifiability Approach to detection of malingered physical symptoms. *Psychology, Crime & Law*, *23*(8), 717–729. https://doi.org/10.1080/1068316X.2017.1302585.

Boskovic, I., Dibbets, P., Bogaard, G., Hope, L., Jelicic, M., & Orthey, R. (2019). Verify the scene, report the symptoms: Testing the Verifiability Approach and SRSI in the detection of fabricated PTSD claims. *Legal and Criminological Psychology*, *24*, 241–257. https://doi.org/10.1111/lcrp.12149.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., … Wu, H. (2018). Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nature Human Behaviour*, *2*, 637–644. https://doi.org/10.1038/s41562-018-0399-z.

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

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers*, *28*, 1–11. https://doi.org/10.3758/BF03203630.

Harvey, A. C., Vrij, A., Leal, S., Lafferty, M., & Nahari, G. (2017). Insurance based lie detection: Enhancing the verifiability approach with a model statement component. *Acta Psychologica*, *174*, 1–8. https://doi.org/10.1016/j.actpsy.2017.01.001.

Honts, C. R. (2014). Countermeasures and credibility assessment. Raskin D. C., Honts C. R., Kircher J. C, In *Credibility assessment: Scientific research and applications*. Cambridge, MA: Academic Press; Retrieved from https://doi.org/10.1016/B978-0-12-394433-7.00004-X.

JASP Team (2019). JASP (Version 0.12.2.0) [Computer software].

Jefferys, W. H., & Berger, J. O. (1992). Ockham's razor and Bayesian analysis. *American Scientist*, *80*, 64–72.

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. *Psychological Review*, *88*, 67–85. https://doi.org/10.1037/0033-295X.88.1.67.

Jupe, L. M., Leal, S., Vrij, A., & Nahari, G. (2017). Applying the Verifiability Approach in an international airport setting. *Psychology, Crime & Law*, *23*(8), 812–825. https://doi.org/10.1080/1068316X.2017.1327584.

Lange, F. (2020). Are difficult-to-study populations too difficult to study in a reliable way ? Lessons learned from meta-analyses in clinical neuropsychology. *European Psychologist*, *25*(1), 41–50. https://doi.org/10.1027/1016-9040/a000384.

Luke, T. J. (2018, June). *Lessons from Pinocchio: Cues to deception may be highly exaggerated*, pp. 1–30. Retrieved from https://doi.org/10.17605/OSF.IO/XT8FQ

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*, 99–122. https://doi.org/10.1080/10683160410001726356.

Nahari, G., & Vrij, A. (2019). The verifiability approach: Advances, challenges, and future prospects. Bull R., Blandón-Gitlin I, In *The Routledge*

international handbook of legal and investigative psychology. Oxfordshire: Taylor & Francis Ltd; Retrieved from https://doi.org/10.4324/9780429326530-15.

Nahari, G., Vrij, A., & Fisher, R. P. (2014a). Exploiting liars' verbal strategies by examining the verifiability of details. *Legal and Criminological Psychology*, *19*, 227–239. https://doi.org/10.1111/j.2044-8333.2012.02069.x.

Nahari, G., Vrij, A., & Fisher, R. P. (2014b). The verifiability approach: Countermeasures facilitate its ability to discriminate between truths and lies. *Applied Cognitive Psychology*, *28*, 122–128. https://doi.org/10.1002/acp.2974.

Open Science Collaboration. (2015). PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science (New York, N.Y.)*. *349*(6251): aac4716. https://doi.org/10.1126/science.aac4716.

Revelle, W. (2015). *Package "psych" - Procedures for Psychological, Psychometric and Personality Research*. R Package.

Revelle, W. (2012). *An introduction to psychometric theory with applications in R*. The Personality Project. Retrieved from https://doi.org/10.2307/2713408.

Singh, J. P., Grann, M., & Fazel, S. (2013). Authorship bias in violence risk assessment? A systematic review and meta-analysis. *PLoS One*, *8*, e72484. https://doi.org/10.1371/journal.pone.0072484.

Vrij, A., Meissner, C. A., Fisher, R. P., Kassin, S. M., Morgan, C. A., & Kleinman, S. M. (2017). Psychological perspectives on interrogation. *Perspectives on Psychological Science*, *12*, 927–955. https://doi.org/10.1177/1745691617706515.

Vrij, A., & Nahari, G. (2019). The verifiability approach. Dickinson J. J, Schreiber Compo N., Carol R., Schwartz B. L., McCauley M., In *Evidence-based investigative interviewing*. London: Routledge; Retrieved from https://doi.org/10.4324/9781315160276-7.