# Applying the Verifiability Approach to Deception Detection in Alibi Witness Situations

Zarah Vernham[1]

Aldert Vrij

Galit Nahari

Sharon Leal

Samantha Mann

Liam Satchell

Robin Orthey

[1] Correspondence concerning this article should be addressed to: Dr Zarah Vernham, University of Portsmouth, Department of Psychology, King Henry Building, King Henry 1 Street, Portsmouth, PO1 2DY, UK, or via e-mail: zarah.vernham@port.ac.uk.

**Names and contact information of all authors:**

1. Dr Zarah Vernham: University of Portsmouth, Department of Psychology, King Henry Building, King Henry 1 Street, Portsmouth, PO1 2DY, UK. E-mail: zarah.vernham@port.ac.uk.
2. Professor Aldert Vrij: University of Portsmouth, Department of Psychology, King Henry Building, King Henry 1 Street, Portsmouth, PO1 2DY, UK. E-mail: aldert.vrij@port.ac.uk.
3. Professor Galit Nahari: Bar-Ilan University, Department of Criminology, Building 213, Israel. E-mail: naharigalit@gmail.com.
4. Dr Sharon Leal: University of Portsmouth, Department of Psychology, King Henry Building, King Henry 1 Street, Portsmouth, PO1 2DY, UK. E-mail: sharon.leal@port.ac.uk.
5. Dr Samantha Mann: University of Portsmouth, Department of Psychology, King Henry Building, King Henry 1 Street, Portsmouth, PO1 2DY, UK. E-mail: samantha.mann@port.ac.uk.
6. Dr Liam Satchell: University of Winchester, Department of Psychology, Winchester, Hampshire, SO22 4NR. E-mail: Liam.Satchell@winchester.ac.uk.
7. Dr Robin Orthey: Kwansei Gakuin University, Nishinomiya, Japan. E-mail: dmh97124@kwansei.ac.jp.

**Abstract**

The application of alibi witness scenarios to deception detection has been overlooked. Experiment 1 was a study of the verifiability approach in which truth-telling pairs completed a mission together, whereas in lying pairs one individual completed this mission alone and the other individual committed a mock theft. All pairs were instructed to convince the interviewer that they completed the mission together by writing individual statements on their own followed by a collective statement together as a pair. In the individual statements, truth-telling pairs provided more checkable details that demonstrated they completed the mission together than lying pairs, whereas lying pairs provided more uncheckable details than truth-telling pairs. The collective statements made truth-telling pairs provide significantly more checkable details that demonstrated they were together in comparison to the individual statements, whereas no effect was obtained for lying pairs. Receiver Operating Characteristic curves revealed high accuracy rates for discriminating between truths and lies using the verifiability approach across all statement types. Experiment 2 was a lie detection study whereby observers' abilities to discriminate between truths and lies using the verifiability approach were examined. This revealed that applying the verifiability approach to collective statements improved observers' ability to accurately detect deceit. We suggest that the verifiability approach could be used as a lie detection technique and that law enforcement policies should consider implementing collective interviewing.

*Keywords:* verifiability approach, alibi witness, collective interviewing, lie detection, consistency.

**Applying the Verifiability Approach to Deception Detection in Alibi Witness Situations**

An alibi witness (often referred to as a person corroborator) is defined as someone who can provide an account of the whereabouts of a suspect at a location other than the crime scene at the time the crime took place (Burke, Turtle, & Olsen, 2007; Dahl & Price, 2012). Alibi witnesses are frequently used by defendants in court (Burke & Turtle, 2003). However, one problem with interpreting alibi witness evidence is that it can sometimes be false. Given that 61% of people believed they could find a false alibi witness to corroborate their story (Culhane, Hosch, & Kehn, 2008), that 82% of people reported they would lie for a romantic partner, and that 68% reported they would lie for their oldest/best friend (Hosch, Culhane, Jolly, Chavez, & Shaw, 2011), false alibi witnesses are likely to be common.

Scenarios involving alibi witnesses differ from those that do not because there is not only a suspect to question, but also an alibi witness. The alibi witness is likely to be telling the truth about his/her whereabouts and activities, but lying about being with the suspect. In such a case, the alibi witness is telling an embedded lie (see Vrij, Granhag & Porter, 2010). Such lies are difficult to uncover and require specific lie detection techniques, such as the verifiability approach (Nahari & Vrij, 2014).

Despite investigators often having to determine whether the evidence provided by an alibi witness is true or false (Culhane et al., 2013), little deception research has explored investigations involving an alibi witness and how one can tell whether the alibi witness evidence is true or false. We carried out two experiments. The first experiment applies the verifiability approach to alibi witness scenarios and the second experiment explores whether

observers (i.e. lay persons whom make judgements about the veracity status of statements) can accurately apply the verifiability approach to correctly classify the statements from alibi witnesses as true or false.

**Verifiability Approach**

The Verifiability Approach was developed by Nahari, Vrij and Fisher (2014a). This approach is based on two assumptions that result in a dilemma for liars. First, research has frequently demonstrated that more detailed accounts signify truthfulness (see Amado, Arce, Fariña, & Vilariño, 2016 for a review) and, as a result, liars want to provide numerous details in order to make an honest impression (Nahari, Vrij, & Fisher, 2012). Second, liars are simultaneously motivated to avoid mentioning details that can be checked and result in the investigator uncovering their lies (Nahari et al., 2012). To accommodate these conflicting aims, liars employ a strategy that focuses on including details that cannot be checked (referred to as unverifiable or uncheckable details, e.g. "I saw some joggers in the park"), and avoid including details that can be checked (referred to as verifiable or checkable details, e.g. "As I entered the park at 9:15am, I bumped into my friend, George, from Rugby"; Nahari, Vrij, & Fisher, 2014a, 2014b).

The current study focuses on the special case in which pairs of suspects are interrogated together, so-called collective interviewing. Collective interviewing can be used in addition to individual interviewing and enables social indicators of deceit to be examined (Vernham & Vrij, 2015), that is, cues to deceit associated with how the group members interact and communicate with one another (Driskell, Salas, & Driskell, 2012; Vernham, Vrij, Mann, Leal, & Hillman, 2014; Vrij et al., 2012). Currently, police typically conduct individual interviews during investigations regardless of the number of interviewees to be questioned. However, collective interviewing is implemented in some existing procedures.

For example, in the United Kingdom, immigration officers occasionally use collective interviewing after individual interviewing when attempting to uncover sham marriages (Home Office, 2013), and in Israel, police detectives sometimes interview multiple suspects collectively if they have provided contradicting versions of events during their individual interviews. Finally, collective interviews are often carried out following individual interviews within some International airports if an individual travelling as part of a group, or if a whole group of individuals, raise suspicion. By extending the research agenda to include collective interviewing, we can uncover new applied contexts where collective interviewing may be appropriate and can therefore inform on best practice (see Vernham & Vrij, 2015 for a review of collective interviewing). For the purpose of the current study, verifiability is defined as any detail that proves the pair (i.e. the suspect and the alibi witness) were together at a location other than the crime scene at the time the crime took place. Hence, the alibi witness merely saying s/he was with the suspect at the time of the crime does not count as verifiability. However, if the alibi witness mentions a third aspect to them being together (e.g. another person or CCTV) then this would count as verifiability.

Nahari and Vrij (2014) applied the verifiability approach to pairs by considering the case of alibi witnesses. It was found that 88% of the pairs could be correctly classified by the verifiability approach. The current study adds to the work of Nahari and Vrij (2014) in several ways: First, the participants in Nahari and Vrij (2014) only had to write collective statements, whereas the participants in the first experiment of the current study were required to write both an individual and a collective statement. This reflects real life better: When collective interviewing is implemented in practice, it is generally used as a follow-up to individual interviewing. Second, Nahari and Vrij (2014) only measured checkable details that demonstrated the pair were together. The current study had three categories of verifiable

details: (1) Checkable details that demonstrate the pair were together (i.e. details that can be verified by an investigator and prove the pair were together at the time the crime was committed), (2) Checkable details that do not demonstrate the pair were together (i.e. details that can be verified by an investigator, but only prove that one member of the pair was at a location other than the crime scene at the time the crime was committed, not both members of the pair), and (3) Uncheckable details (i.e. details that cannot be verified during an investigation). This division of verifiable details into three categories is important because (i) the first category allows for the replication of the findings obtained by Nahari and Vrij (2014); (ii) the second category is more applicable to alibi witness research because it allows us to take into account the fact that even lying alibi witnesses might provide checkable details that demonstrate their activities, but not necessarily that they carried them out with their partner; and (iii) the third category allows us to explore whether liars compensate for the lack of reporting verifiable details by reporting more unverifiable details, which also enables us to examine the proportion of the checkable details in the statement(s), as a within-subjects measurement. Recently, researchers have started to discuss the importance of within-subjects measures in verbal lie detection (Nahari & Vrij, 2014, 2015; Vrij, 2016). The key argument is that verbal cues that appear as highly diagnostic in verbal deception research where veracity differences are assessed at a between-subjects (i.e. group) level (e.g. the number of details liars provide compared to truth-tellers; Amado, Arce, & Fariña, 2015) are of little value in real life contexts where assessments are made at a within-subjects (i.e. individual) level, due to large individual differences in the reporting of such cues (e.g. the amount of details someone typically provides). Thus, verbal cues (e.g. proportion of checkable details) need to be developed that control for these individual differences in providing details (Vrij,

2016). Finally, unlike the current study, Nahari and Vrij (2014) did not incorporate an additional experiment to test the validity of the verifiability approach as a lie detection tool.

**Experiment 1: Hypotheses**

It is hypothesised that truth-telling pairs will provide significantly more checkable details that demonstrate the pair were together in both the individual and collective statements than lying pairs (Hypothesis 1). Although both truth-telling pairs and lying pairs are expected to provide many uncheckable details, it is predicted that when the proportion of uncheckable details is calculated (i.e. the total number of uncheckable details divided by the total number of checkable and uncheckable details), lying pairs will provide significantly more uncheckable details than truth-telling pairs in both the individual and collective statements (Hypothesis 2). No difference is expected between pairs of truth-tellers and pairs of liars in terms of checkable details that do not demonstrate the pair were together because both truth-telling pairs and lying pairs will have had at least one member of the pair complete the non-criminal activities. Thus, both truth-telling pairs and lying pairs will be able to show through the provision of checkable detail that at least one member of their pair completed the non-criminal activities.

It is further hypothesised that truth-telling pairs will provide significantly more checkable details that demonstrate they were together in the collective statement compared to the individual statements (Hypothesis 3). This is because, during the collective statement, joint recall is occurring with truth-telling pairs focusing on recollecting shared events (see literature on transactive memory; Hollingshead, 1998; Wegner, 1987) and we believe that this joint recall will reflect itself in the collective statement as checkable details that demonstrate the pair were together. This pattern will emerge significantly less for lying pairs because they

cannot recall actual shared memories leading them to merely repeat details from their individual statements.

## Experiment 1: Method

### Participants

The current research was approved by a Science Faculty Ethics Committee within a U.K. University. A total of 120 participants (30 truth-telling pairs and 30 lying pairs) took part in this first experiment. However, one lying pair was excluded as they did not correctly follow the instructions of the experiment. The mean age of the remaining 118 participants was 24.38 years (*SD* = 10.48), 34 were male and 84 were female. Of the pairs, 32 were female pairs, 7 were male pairs, and 20 were mixed gender pairs.

### Design

This first experiment used a mixed design with Veracity (truth versus lie) as the only between-subjects factor and Statement (individual versus collective) as the only within-subjects factor. The proportion of *checkable details (proof pair together), checkable details (other),* and *uncheckable details* were the three dependent variables.

### Procedure

Participants were recruited via online advertisements, the university staff and student portals, and word of mouth. All participants were told prior to signing up to the experiment that it was an experiment investigating the interactions occurring between friends and therefore they were required to sign up in pairs.

Upon arrival at the Psychology department, all pairs of friends were required to read and sign an informed consent form and were randomly assigned to one of the two veracity conditions. They were told by the experimenter that they were going to complete a task together (truth-telling pairs) or complete separate tasks (lying pairs).

Truth-telling pairs were sent on a mission around a nearby park together. The park has many features, such as a children's play area, an animal enclosure, and several war monuments. Truth-telling pairs were provided with instructions of what to do on their mission around the park, a map with directions of how to get there, a map of the park itself, and a task sheet asking seven questions about different areas of the park. Their mission was to go around the park and work together as a pair to answer the seven questions on the task sheet in the order in which they were asked. Despite the experimenter requesting participants to answer the questions in a specific order, approximately 17% of participants mentioned that they did not follow the tasks in chronological order because they accidentally came across an answer to a later question first. The maps provided could be used to help them locate the answers they required. Questions on the task sheet included "How many slides are there in the children's play area?", "Name five animals that live within the enclosure at the centre of the park", and "What is the date on which the Chinese bell monument was captured?". Pairs were instructed to stay together at all times, working together to answer each of the questions. The questions could only be answered correctly if the pair actually went to the specific places within the park, providing some ground truth that the truth-telling pairs did do the entire mission. Although it could be that participants searched on the internet for the answers (e.g. using their mobile phone), they would only have been able to find answers to three of the seven questions. When asked, no participants admitted to having used the internet to obtain any of the answers. On completion of the tasks, the pairs returned to the Psychology department. On returning to the Psychology department, they handed the experimenter the task sheet, which enabled the experimenter to check that they had completed each of the tasks. When back at the department, the experimenter informed the pair that a crime had taken place and that one of them matched the description provided of the person who was

seen leaving the office in which the crime had occurred (the pair member chosen to match the description was picked at random by the experimenter). Therefore, this individual became the 'suspect' (who was innocent) and his or her friend became the 'true alibi witness'.

Lying pairs were separated and randomly assigned a mission. One individual was instructed to do exactly what the truth-telling pairs were asked to do, but on his or her own rather than with his or her friend. The other individual was instructed to commit a mock crime on his or her own. S/he was given a key and required to follow directions to a locked office in the Psychology department. S/he was to unlock the office, steal £20 from a purse on the desk within the office, lock the office and return to the experimenter with the £20. The £20 was returned to the experimenter following completion of the study. S/he was to do this as quickly as possible and without being seen. When both individuals had completed their tasks, they were reunited as a pair and informed that a crime had been reported and that the individual who had actually completed the mock crime matched the description provided of the person who was seen leaving the office in which the crime occurred. Therefore, this individual became the 'suspect' (who was guilty) and his or her friend was instructed be a 'false alibi witness'.

The task given to all pairs was to convince an investigator that they were together at all times when the crime was committed. They were instructed to state that they had been completing a mission around the park together at the time the money was stolen. Therefore, truth-telling pairs (both the 'suspect' and the 'true alibi witness') were required to tell the truth about their whereabouts and activities at the time of the crime. The lying pairs on the other hand were required to lie, with the suspect having to lie entirely about his or her whereabouts and activities, claiming that s/he was with the 'false alibi witness', whereas his or her friend (the 'false alibi witness') had to tell the truth about his or her whereabouts and

activities but lie about being alone when completing the mission – That is, the 'false alibi witness' had to say that s/he completed the mission together with the 'suspect' in order to try to exonerate his or her friend. The 'false alibi witness' was made aware that his or her friend (the 'suspect') had stolen the money and therefore s/he was intentionally lying about being with his or her friend at the time the crime was committed.

All pairs were given as much time as they wanted to prepare for questioning and to get their stories straight. They were told to focus on discussing how they were going to prove their own innocence or the innocence of their partner. All pairs were informed prior to their preparation talks that they would be required to write a statement on their own and then a second statement together as a pair. Therefore, if differences between truth-tellers and liars were to emerge, this would not be because the individual or collective statements took the pairs by surprise.

Once the pairs stated they were ready to be questioned, they were separated and individually completed pre-questioning questionnaires. These were completed to get an understanding of the degree to which pairs had prepared for questioning and whether the preparation discussions differed between truth-telling pairs and lying pairs. The pre-questioning questionnaire asked participants to rate on 7-point Likert scales the thoroughness (from [1] incomplete to [7] thorough), sufficiency (from [1] insufficient to [7] sufficient), quality (from [1] very poor to [7] very good), and usefulness (from [1] pointless to [7] useful) of their preparation discussion. It also asked the participants to rate how much they discussed with their partner about what to say during questioning (ranging from [1] not at all to [7] thoroughly).

Subsequently, each member of the pair separately typed up individual statements on a laptop answering the following question: *'Describe in as much detail as possible what you*

*were doing at the time the crime took place. Think about your whereabouts, your activities, the people you were with, what you saw, what you heard, how you felt etc.'*. The statement system was set up on the laptop to look like the statement was being sent to an investigator. That is, all participants were manipulated to believe that the investigator was receiving their statements once complete. Once both members of the pair had completed their individual statements, they were put together to write a collective statement answering the same question. The pairs could speak freely to one another and could choose who typed up the statement. They were reminded prior to writing the individual statements and prior to writing the collective statement that their task was to convince an investigator that they were together the whole time around the park at the time the crime occurred. They were led to believe that the investigator was receiving their written statements (both individual and then collective) once they clicked submit on the laptop within the statement system.

The participants were not informed about any information the investigator would be looking for in their statements. To motivate participants to perform well during the experiment, they were told that if they were believed by the investigator they would each receive £10. However, if they were not believed they would receive no money and would be required to write a further statement about their whereabouts and activities at the time the crime took place. To ensure that the experiment was ethical and equal for all participants, the experimenter told them at the end of the experiment that the investigator believed they were telling the truth, and so all participants were paid £10 each.

Following participation, a post-questioning questionnaire was completed individually and at this stage all participants were instructed to be truthful about their experience of writing the statements. In this questionnaire, participants were asked to rate on a 7-point Likert scale from [1] not at all motivated to [7] extremely motivated, the extent to which they

felt motivated to appear convincing during questioning. They were also asked to rate their confidence in receiving £10 and their confidence about having to write a further statement (both on 7-point Likert scales from [1] none at all to [7] very likely). Additionally, participants were asked to rate on 7-point Likert scales (ranging from [1] easy to do to [7] difficult to do) the extent to which they found writing their individual statement and the collective statement easy or difficult to do. Finally, to explore how honest participants were in their statements, they rated on scales from 0% to 100% with 10% intervals the extent to which they had told the truth during the individual statement and then the collective statement.

To make sure there were no significant confounding differences between the truth-telling pairs and lying pairs, all pairs were asked, in the post-questioning questionnaire, about their knowledge of the park, how they perceived their level of friendship with their study partner, and how long the pair had been friends for.

Once the post-questioning questionnaire was completed by both members of the pair, they were each given a debriefing form and provided with the opportunity to ask the experimenter questions. The whole study took pairs of participants between 60 minutes and 90 minutes to complete.

**Coding Statements for Verifiability**

The statements were coded by a rater who was blind to the hypotheses and veracity status of the pairs. The three dependent variables (*checkable details (proof pair together), checkable details (other), uncheckable details*) were coded for each of the individual statements and the collective statements separately. The two individual statements of each pair were then compared and duplicates of checkable or uncheckable details between the individual statements were removed allowing for one total score from the two individual

statements to be calculated for each variable. The removal of duplicate details between individual statements of the same pair needed to be done to remove the confound of there being only one collective statement and two individual statements and to prevent repetition between individuals of the same pair from distorting the data. Overall, this meant that each pair obtained two total frequency scores for each of the three verifiable variables, one score from the individual statements and one score from the collective statement.

*Checkable details (proof pair together)* was the number of details provided by the participant(s) that could be verified and demonstrated the pair were together at the time of the crime (e.g. "We bumped into our tutor Anne and spoke to her for a bit" or "There was CCTV in Guildhall square that would have picked us up").

*Checkable details (other)* was the number of details provided by the participant(s) that could be verified but did not necessarily demonstrate that the pair were together at the time of the crime. For example, providing details that show that one member of the pair was at the park but not necessarily the other member of the pair (e.g. "The park warden was feeding the animals whilst I was there…He saw me writing down the animals on my answer sheet" or "I saw a litter picker in the park who questioned me about what I was doing").

*Uncheckable details* was the number of details provided by the participant(s) that could not be verified (e.g. "We spent two whole minutes staring at the guinea pigs before moving on" or "There were no children in the play area when we walked past").

Each type of detail was converted into a proportion variable by dividing each type of verifiable detail by the total number of all three types of verifiable details. For example: Proportion of checkable details (proof pair together) = total checkable details (proof pair together) / {total checkable details (proof pair together) + total checkable details (other) + total uncheckable details}. A relative measurement was used because we wanted to take into

account individual differences in the richness of details (e.g. Nahari & Pazulo [2015] found gender differences in the richness of detail when telling the truth). Also, truth-telling pairs ($M$ = 209.95, $SD$ = 120.93) provided significantly more details overall in their statements compared to lying pairs ($M$ = 153.65, $SD$ = 58.79), $t(116) = 3.20$, $p = .002$, $d = 0.77$. For this reason, we needed to use a within-subjects measure (i.e. the proportion of (un)verifiable details).

A second coder, also blind to the hypotheses and veracity status of the pairs, coded the individual and collective statements obtained from 16 pairs for the total number of times each of the three verifiable variables occurred. Intra-class correlation coefficients (ICCs) were then calculated between the two individual raters. The inter-rater reliability between the two coders was very good for both the individual and collective statements with each of the ICCs demonstrating strong agreement between the two raters (checkable details (proof pair together): ICCs = .82 and .91; checkable details (other): ICCs = .88 and .89; and uncheckable details: ICCs = .90 and .82).

**Analysis of Data**

Bayesian analyses, using JASP software, were conducted to supplement all statistical analyses, with the default Cauchy's prior of .707 used for the Bayesian *t*-tests (Lakens, 2016). Therefore, within the statistical analyses, we report not only the Cohen *d*'s as a measure of effect size but also the Bayes factor, $BF_{10,}$ for main effects. Bayes factors (BFs) enable us to quantify the evidence for the alternative hypothesis (presence of an effect) relative to the null hypothesis (absence of an effect), and vice versa. That is, as $BF_{10}$ increases, there is more evidence in support of the alternative hypothesis, but the inverse yields the opposite (i.e. 1/ $BF_{10}$) and provides more evidence in support of the null hypothesis (see Jaroz & Wiley, 2014). In line with the cut-off points outlined by Jeffreys (1961), BFs between 1 and 3

suggest weak evidence, BFs between 3 and 10 suggest strong evidence, and BFs > 10 indicate very strong evidence for the alternative hypothesis relative to the null hypothesis. If we obtain evidentially weak Bayes factors suggesting an absence of evidence, then the results can be judged as anecdotal or inconclusive (Lee & Wagenmakers, 2013).

Receiver Operating Characteristic (ROC; Mossman, 1994) curves were created to evaluate the efficiency of the verifiability approach for detecting deceit. ROC curves reflect the degree of separation between the distributions of the proportion of checkable to uncheckable details and the detection of truth-telling versus lying participants. That is, a ROC curve plots the rate of true positives against the rate of false positives for the proportion of (un)checkable details and represents the trade off in specificity that occurs as sensitivity increases. The Area Under the Curve (AUC) of the ROC graph is the index for interpreting the overall predictive validity of the verifiability approach. AUC values of 0.5 correspond to no-better-than chance prediction and values of 1.0 correspond to perfect efficiency. In general, an AUC of 0.5 is therefore a lack of ability to discriminate between truths and lies, 0.7 to 0.8 suggests acceptable discrimination, 0.8 to 0.9 suggests excellent discrimination, and more than 0.9 suggests outstanding discrimination (Hosmer & Lemeshow, 2000).

## Experiment 1: Results

### Pre-questioning Questionnaire

Preparation time was offered to all participants. However, only six truth-telling pairs compared to all 29 lying pairs chose to prepare prior to writing their statements. This finding is frequently obtained in deception detection studies (e.g. Vernham et al., 2014; Vrij et al., 2009) and is not surprising: Truth-telling pairs merely rely on memory to recall events, something which lying pairs cannot do because the events did not actually occur. Hence, lying pairs rely on a fabricated story that they need to plan and collaborate on, in order to get

their story straight, keep details simple, and avoid contradictions (Vrij, Mann, Leal, & Granhag, 2010). Of those pairs who did choose to prepare, the time spent preparing ranged from 1.02 minutes to 19.02 minutes. A *t*-test revealed that lying pairs ($M$ = 7.90 mins, $SD$ = 3.84) spent significantly longer preparing than truth-telling pairs ($M$ = 2.14 mins, $SD$ = 1.39), $t(33) = 3.59$, $p = .001$, $d = 1.99$, $BF_{10} = 6.137^{e+11}$. This finding is not surprising either: It simply reflects the fact that liars have more work to do to get their story straight than truth-tellers.

A one-way (Veracity: truth vs. lie) between-subjects MANOVA was conducted to examine if there were any significant differences between truth-tellers and liars in terms of (i) how they rated their preparation discussion prior to writing their statements and (ii) how much they discussed with their partner about what to include in their written statements. The MANOVA revealed a significant multivariate main effect for Veracity, Wilks' $\lambda = .74$, $F(5, 62) = 4.39$, $p = .002$, $\eta_p^2 = .26$. Significant univariate main effects for Veracity were obtained with liars rating the preparation discussion as significantly more useful and more sufficient than truth-tellers. Liars also reported discussing with their partner about what to write in their statements significantly more thoroughly than truth-tellers. No significant differences were found between truth-tellers and liars in terms of how they rated their preparation discussion for thoroughness or quality (see Table 1).

INSERT TABLE 1 HERE

**Post-questioning Questionnaire: Motivation, Manipulation Checks, and Writing Statements**

The vast majority of participants self-reported that they were motivated to appear convincing during the interview, with 85.6% of the sample scoring 5 or higher on the 7-point Likert scale.

A one-way (Veracity: truth vs. lie) between-subjects MANOVA was conducted to investigate whether there were any significant differences between truth-tellers and liars in terms of their self-reporting of motivation, confidence, how difficult they found writing the individual and collective statements, and how much they told the truth on both the individual and collective statements. The MANOVA revealed significant multivariate main effects for Veracity, Wilks' $\lambda = .07$, $F(7, 108) = 214.09$, $p < .001$, $\eta_p^2 = .93$. As shown in Table 1, significant univariate main effects for Veracity were obtained with liars reporting significantly more motivation to appear convincing than truth-tellers. In terms of confidence, truth-tellers reported that they were more confident than liars that they would receive £10, whereas liars were more confident than truth-tellers that they would have to write a further statement. Furthermore, liars found writing both the individual statement and the collective statement significantly more difficult than truth-tellers. Finally, on both the individual statement and collective statement, truth-tellers reported staying closer to the truth than liars (see Table 1).

To ensure that any findings obtained in the current experiment were the result of Veracity and not the result of participants in one condition knowing the park better than participants in the other condition, an independent samples *t*-test was conducted to compare whether there was a significant difference between truth-tellers and liars in regard to their self-reported knowledge of the park. No significant difference was found ($p = .638$, $BF_{10} = 0.27$). Additionally, to ensure that any findings obtained were not confounded by the level of friendship of each pair, a one-way MANOVA was conducted on the participants' self-report data about how friendly they were with their study partner. The MANOVA indicated that there were no significant differences between truth-telling pairs and lying pairs in terms of how they rated their friendship on four 7-point Likert scales that measured; (i) labelling (from

[1] strangers to [7] best friends); (ii) closeness (from [1] distant to [7] intimate); (iii) importance (from [1] unimportant to [7] important); and (iv) trustworthiness (from [1] distrusting to [7] trusting) (means for truth-tellers ranged from 5.83 to 6.63; means for liars ranged from 5.55 to 6.40; *p*-values ranged from .086 to .218, $BF_{10's}$ ranged from 0.39 to 0.75). Finally, to ensure that any findings obtained were not confounded by the self-reported length of friendship of each pair, an independent samples *t*-test was conducted. This revealed that there was no significant difference between pairs of truth-tellers and pairs of liars in terms of friendship length ($p = .094$, $BF_{10} = 0.88$).

**Hypotheses Testing: Proportion of (Un)Checkable Details**

Three two-way mixed ANOVAs were conducted on the pair level. Since the dependent variables of these analyses were proportion measures, they were inter-dependent. To deal with this issue, Bonferroni correction was applied ($\alpha = .05 / 3$). Hence, each analysis was tested using a required significance level of .017.

The first 2 x 2 mixed design ANOVA was conducted with Veracity (truth vs. lie) as the between-subjects factor, Statement (individual vs. collective) as the within-subjects factor, and *proportion of checkable details (proof pair together)* as the dependent variable. The ANOVA revealed that truth-telling pairs ($M = .32$, $SD = .14$, 95% CI [.29, .38]) provided a significantly higher proportion of checkable details (proof pair together) than lying pairs ($M = .15$, $SD = .11$, 95% CI [.10, .20]), $F(1, 57) = 28.25$, $p < .001$, $\eta_p^2 = .33$, $d = 1.35$, 95% CI [0.71, 1.83], and a $BF_{10}$ of 10760.76 provided very strong evidence for this effect. Furthermore, a significantly higher proportion of checkable details (proof pair together) were provided in the collective statement ($M = .28$, $SD = .23$, 95% CI [.23, .33]) compared to the individual statement ($M = .20$, $SD = .15$, 95% CI [.17, .24]), $F(1, 57) = 10.32$, $p = .002$, $\eta_p^2 = .15$, $d = 0.41$, 95% CI [0.02, 0.75], and a $BF_{10}$ of 20.12 again provided very strong support for

this effect. Finally, a significant Statement X Veracity interaction effect was obtained for checkable details (proof pair together), $F(1, 57) = 7.20$, $p = .010$, $\eta_p^2 = .11$. A simple main effects analysis demonstrated that the collective statement led to a significantly higher proportion of checkable details (proof pair together) than the individual statements for pairs of truth-tellers, but no such difference was found between the individual statements and the collective for pairs of liars. The Bayes factor analyses for these simple main effects showed very strong support for the truth-telling pairs, whilst also showing weak, inconclusive evidence for the absence of a simple main effect in lying pairs (see Table 2). These findings provide support for Hypotheses 1 and 3.

INSERT TABLE 2 HERE

The second 2 x 2 mixed design ANOVA was conducted with Veracity (truth vs. lie) as the between-subjects factor, Statement (individual vs. collective) as the within-subjects factor, and *proportion of checkable details (other)* as the dependent variable. The ANOVA revealed that there was no significant difference between truth-telling pairs and lying pairs in terms of the proportion of checkable details (other) provided ($p = .124$, $BF_{10} = 0.51$). There was also no significant difference between the individual and collective statements in terms of proportion of checkable details (other) ($p = .082$, $BF_{10} = 0.61$) nor was there a significant Statement X Veracity interaction effect ($p = .536$). The Bayes factor analyses for checkable details (other) suggest weak, insufficient evidence for the absence of both Veracity and Statement main effects.

The third 2 x 2 mixed design ANOVA was conducted with Veracity (truth vs. lie) as the between-subjects factor, Statement (individual vs. collective) as the within-subjects factor, and *proportion of uncheckable details* as the dependent variable. The ANOVA revealed that lying pairs ($M = .59$, $SD = .20$, 95% CI [.52, .65]) provided a significantly

higher proportion of uncheckable details compared to truth-telling pairs ($M$ = .36, $SD$ = .10, 95% CI [.29, .42]), $F(1, 57) = 26.12$, $p < .001$, $\eta_p^2 = .31$, $d = 1.46$, 95% CI [0.81, 1.94], $BF_{10} = 33138.18$. There was no significant difference between the individual and collective statements in terms of proportion of uncheckable details ($p = .087$, $BF_{10} = 0.60$) nor was there a significant Statement X Veracity interaction effect ($p = .217$). The Bayes factor analyses for uncheckable details showed very strong evidence for the Veracity main effect as well as showing weak, inconclusive evidence for the absence of the Statement main effect. These findings support Hypothesis 2.

**Receiver Operating Characteristics (ROC)**

As no significant findings were obtained for checkable details (other), the ROC curves were conducted for the proportion of checkable (proof pair together) to uncheckable details only (i.e. total checkable details (proof pair together) divided by total uncheckable details). The ROC curves were first conducted regardless of statement type and then depending on statement type (i.e. individual vs. collective). All ROC curves were significant at $p < .001$ with AUC values ranging from .789 to .895 (see Table 3). Hence, the proportion of checkable to uncheckable details was predictive of veracity across all statement types, which provides strong evidence in support of the efficiency of the verifiability approach as a systematic lie detection tool that can accurately discriminate between pairs of liars and pairs of truth-tellers.

## Experiment 2: Discriminating between Truth-Telling Pairs and Lying Pairs

In support of Hypotheses 1 and 2, truth-telling pairs, compared to lying pairs, included a higher proportion of checkable details that demonstrated the pair were together in both the individual and collective statements, whereas, lying pairs, compared to truth-telling pairs, included a higher proportion of uncheckable details in both the individual and

collective statements. The proportion of checkable details that demonstrated the pair were together increased between the individual and collective statements for truth-telling pairs but not for lying pairs, supporting Hypothesis 3. Additionally, ROC analyses demonstrated the high potential of the verifiability approach for correctly classifying pairs of participants based on veracity.

The fact that the verifiability approach revealed cues to deceit does not automatically mean that observers will be able to accurately discriminate between pairs of truth-tellers and pairs of liars when informed about the verifiability approach. ROC curves, whilst objective, do not take into account human bias and subjectivity when predicting the veracity status of the participants. This second experiment therefore investigated; (i) whether the verifiability approach could be used by observers to make accurate veracity judgements regarding single cases; (ii) whether observers could apply the verifiability approach quickly and easily; (iii) whether making the veracity judgement was uninfluenced by other cues (e.g. consistency); and (iv) whether the ability of observers to correctly classify pairs based on veracity was dependent upon whether they read individual statements, a collective statement, or a combination of individual and collective statements. This second experiment is therefore a relevant addition to the previous experiment as it could be argued that the verifiability approach is more valuable if observers are able to appropriately apply the approach and identify (un)checkable details within statements.

We expect that observers will find it easier to apply the verifiability approach when reading a collective statement compared to when reading individual statements or a combination of individual and collective statements. When observers have access to multiple statements they automatically rely on verbal consistency between statements (even when instructed not to do so), interpreting inconsistent statements as deceitful and consistent

statements as truthful (Granhag & Strömwall, 2000; Strömwall, Granhag, & Jonsson, 2003). However, despite consistency being the strongest cue that observers pay attention to when attempting to detect deceit (Potter & Brewer, 1999; Strömwall & Granhag, 2003; Vrij, 2008), research shows that both lay people and legal professionals utilise the consistency cue incorrectly. That is, contrary to popular belief, lying pairs, although more vague, actually appear as consistent if not more consistent than truth-telling pairs. This is because lying pairs provide an alibi and pre-plan their responses (Granhag, Strömwall & Jonsson, 2003; Strömwall, et al., 2003; Vrij, Mann, et al., 2010), whereas truth-telling pairs rely on memory, which is naturally reconstructive in nature (Bartlett, 1932; Loftus, 2005). This relationship between consistency and deception has been illustrated within the 'repeat vs. reconstruct' hypothesis proposed by Granhag and Strömwall (1999): Liars will merely repeat what they have previously said, whereas truth-tellers will change, add and remove information. Hence, consistency may in fact be a sign of lying as opposed to truth-telling (Vredeveldt, van Koppen, & Granhag, 2014).

**Experiment 2: Hypothesis**

It is hypothesised that observers being informed about the verifiability approach will discriminate significantly better between pairs of truth-tellers and pairs of liars when judging veracity based on one collective statement (where consistency is not relevant), than when judging veracity based on two individual statements or a combination of individual and collective statements (where consistency is relevant) (Hypothesis 4).

## Experiment 2: Method

**Participants**

A total of 57 observers with a mean age of 39.81 years (*SD* = 15.36) took part in this experiment, 25 were male and 32 were female. All observers who took part in this lie

detection experiment were volunteers who were not compensated for participating. These volunteers were recruited using an opportunity sample based on who was available and willing to take part. Hence, participants included members of the public, family and friends, and University staff and students who were recruited via word of mouth or in response to an advertisement. Additionally, all observers who took part in this second experiment had not partaken as part of any pair in the first experiment.

**Design**

This second experiment was a one-way repeated measures design with Statement Type as the within-subjects factor. This factor consisted of three levels; individual (observer makes a veracity judgement having read both individual statements from the pair) versus collective (observer makes a veracity judgement having read only the collective statement from the pair) versus combined (observer makes a veracity judgement having read both individual statements and the collective statement from the pair). The statements being judged in this second experiment were all those produced from the pairs of participants within the first experiment (excluding the two lots of statements used as examples; *see Procedure section below*). All 57 observers made a veracity judgement three times, but no observers judged the same pair of participants twice (i.e. the individual statements judged by each observer were written by a different pair to the collective statement, which was again written by a different pair to the combined statements). The order in which each observer read each statement was counterbalanced as was the number of truth-telling and lying pairs that each observer judged (i.e. 27 of the observers judged two lying pairs and one truth-telling pair and 30 of the observers judged two truth-telling pairs and one lying pair). However, no observer had only truth-telling pairs or only lying pairs. The dependent variable for the current experiment was the veracity judgements (accuracy rates).

**Procedure**

Observers were recruited using an opportunity sample and asked to read an information sheet about the experiment. If they were willing to participate, they then signed an informed consent form. Next, observers were provided with (i) information about the procedure of the first experiment (i.e. pairs of participants either completed a mission together [truth-telling pairs] or separately [lying pairs]), (ii) a definition of 'alibi witness', and (iii) their task instructions for this second experiment (i.e. to read a variety of statements and make a judgement about the veracity status of the pair who wrote the statements).

Following this, observers were informed about the two different types of statements they would be judging; individual statements (pair members are separated to write a statement alone) and collective statements (pair members write a statement together). Observers were also taught about the verifiability approach and provided with definitions and examples of checkable and uncheckable details. Checkable details (proof pair together) were defined to observers as details that could be verified by an investigator to demonstrate that the pair were together at a location other than the crime scene at the time the crime took place (e.g. "we walked into the park at 2:15pm and bumped into our friend George from rugby"), and uncheckable details were defined as details that could not be verified by investigators to show the pair were together (e.g. "leaves were falling off the trees in the park"). Observers were told that previous research has found that truth-tellers provide significantly more checkable details than liars, whereas liars provide significantly more uncheckable details than truth-tellers[i]. Finally, participants were shown examples of individual and collective statements from one truth-telling pair and one lying pair with the checkable and uncheckable details highlighted within the statements. These example statements were then removed from being judged by observers in the actual experiment. The remaining 171 statements (114

individual statements and 57 collective statements) from Experiment 1 were then judged by observers in this second experiment. Observers were told that they should use the verifiability approach to make a decision about the veracity status of the statements they will be reading.

Once observers self-reported to understand the verifiable approach (i.e. observers merely told the experimenter that they understood the approach and what they had to do), they completed a few demographic details and read the statement(s) from their first allocated pair. They then completed a questionnaire and judged the veracity status of this first pair. This questionnaire required observers to state whether they thought the pair who wrote the statement(s) were lying or telling the truth. Observers then freely reported (via an open-ended question) the cues which had helped them make their veracity judgement, and rated on a 7-point Likert scale how difficult they found having that particular type of statement (i.e. two individual statements or one collective statement or all three statements [two individual and one collective]) to determine the veracity status of the pair (from [1] very easy to [7] very difficult). Next, observers judged the statement(s) of another pair and again completed the same questionnaire. Finally, observers judged the statement(s) of a third pair and again completed the same questionnaire. To end the experiment, observers were debriefed and provided with the opportunity to ask questions. Participation lasted approximately 30 minutes.

**Subjective Coding: Consistency**

To take into account previous research (e.g. Strömwall & Granhag, 2003; Strömwall, et al., 2003; Vredeveldt et al., 2014) and the belief that consistency may have impacted upon the veracity judgements made by observers when rating multiple statements from the same pair, two coders (blind to the hypotheses and veracity status of the pairs) rated the two

individual statements of each pair and the combined statements (two individual and one collective) of each pair for consistency (i.e. how similar each statement was to one another). Thus, each pair obtained two consistency ratings from each coder on a Likert scale from [1] Not at all consistent to [7] Completely consistent. ICCs were calculated between the two individual coders. The inter-rater reliability between the two coders was good when rating the consistency of the individual statements (ICC = .86) and the combined statements (ICC = .87), demonstrating strong agreement between the two coders. The ratings obtained from the two coders were then averaged to obtain an averaged consistency rating for the individual statements of each pair and an averaged consistency rating for the combined statements of each pair. These mean ratings were then used in future analyses.

**Analysis of Data**

The cues reported to have been used by the 57 observers were coded and computed per Statement Type. A total of 13 different cues were spontaneously mentioned by observers: Verifiability (observer mentions details that can or cannot be checked), number of details (observer mentions a small or large quantity of details), consistency (observer mentions repetitions, contradictions, omissions or commissions), pronouns (observer mentions use of 'I' or 'we'), plausibility (observer states that the information was realistic or believable), unnecessary details (observer mentions information being irrelevant or not needed), feelings/opinions (observer mentions feelings or emotions, such as the pair seemed happy and comfortable with one another, or mentions thoughts or opinions being reported by the pair), equality (observer mentions either that the pair spoke equally and provided the same amount of information or that one member of the pair provided more information than the other member), interactions (observer mentions information about the reporting of joint experiences or activities that seem to have been carried out together), statement length

(observer states that the statement(s) were long or short), overcomplicated information (observer mentions details that made the statement too confusing or complicated), specificity of information (observer mentions specific details or details that were too general/vague), and memory recall (observer mentions how people remember or recall information, or the fact that it is normal for some information to be forgotten and written differently by different people). Each cue could only be mentioned once by each observer. To measure the reliability of the coding, a second rater coded the cues reported by 15 observers across the three types of statements. Inter-rater reliability analyses, using the Kappa statistic, revealed excellent agreement between the two raters in allocating the cues to the 13 categories across each of the three statement types (individual statements: Kappa = .91, $p < .001$; collective statements: Kappa = .90, $p < .001$; combined statements: Kappa = .82, $p < .001$). Manipulation checks were conducted to explore what cues observers reported to be using to make their veracity judgements across the three types of statements. Additionally, both truth accuracy (percentage of true statements that were correctly classified) and lie accuracy (percentage of deceptive statements that were correctly classified) were measured for all 57 observers across the three statement types by giving the observer a 1 if their veracity judgement was correct and a 0 if their veracity judgement was incorrect.

**Lens modelling.** Observers' self-reported use of cues can only inform us about what cues they actively considered when making their judgments. Whilst this can be of interest, we also wished to analyse the degree to which the specific cues of (un)checkable details and consistency influenced observers' judgments. A 'lens modelling' approach to analysis (based on the work of Ego Brunswik; see Karelaia & Hogarth, 2008) allows for the examination of both the presence of cues in the targets and the influence of such cues on observers' judgments. The analysis uses Pearson's *r* correlations to relate the target condition (truth-

telling pair/lying pair), observer judgment (truth/lie), and presence of (un)checkable details or consistency. The correlation between target condition and observer judgement (frequently known as the index of 'achievement' in lens models) can be considered an index of 'accuracy'. The presence (or absence) of this accuracy can then be investigated through the *lens* of (un)checkable details or consistency. For example, if observers are accurate then this may be explained by a significant correlation between the presence of checkable details and the veracity of the target and a correlation (in the same direction) between the presence of checkable details and judgments. If observers are inaccurate in their judgment, a lens model style of analysis can demonstrate if this is due to a lack of relationship between checkable details and target (that is, no correlations between the targets' veracity and the cues) or if this is due to the observers not using checkable details for inference (that is, no correlations between observers' judgments and the cues). Lens models are frequently presented graphically (see Figure 1-5) and have previously been explored as an approach to analysing deception research data (e.g. in a meta-analysis by Hartwig & Bond, 2011). To avoid concerns about multiple comparisons, data will be analysed by considering the size of correlations. Ferguson's (2009) recommended effect sizes will be used as a guide.

## Experiment 2: Results

### Lie Detection: Accuracy

The overall accuracy rate was 56.1% across all three types of statement (truth accuracy = 56.3%, lie accuracy = 56.0%). When judging the individual statements, observers obtained an accuracy rate of 40.3% (truth accuracy = 34.5%, lie accuracy = 46.4%); when judging the collective statements, observers obtained an accuracy rate of 79.0% (truth accuracy = 82.8%, lie accuracy = 75.0%); and when judging the combined statements, observers obtained an accuracy rate of 49.1% (truth accuracy = 51.7%, lie accuracy =

46.4%). A one-way repeated measures ANOVA was conducted with Statement Type (individual vs. collective vs. combined) as the within-subjects factor and accuracy rate as the dependent variable. As with Experiment 1, Bayesian analyses were conducted using the cut-off points outlined by Jeffreys (1961) to complement the statistical analyses. The ANOVA revealed a significant univariate main effect for Statement Type, $F(2, 112) = 10.32$, $p < .001$, $\eta_p^2 = .16$. The total accuracy rate of the collective statements ($M = .79$, $SD = .41$, 95% CI [.68, .90]) was significantly higher than the total accuracy rate of the individual statements ($M = .40$, $SD = .49$, 95% CI [.27, .53]), $F(1, 56) = 17.29$, $p < .001$, $\eta_p^2 = .24$, $d = 0.86$, 95% CI [0.43, 1.19], $BF_{10} = 400.41$, and the total accuracy rate of the combined statements, ($M = .49$, $SD = .50$, 95% CI [.36, .63]), $F(1, 56) = 12.95$, $p = .001$, $\eta_p^2 = .19$, $d = 0.66$, 95% CI [0.24, 0.99], $BF_{10} = 77.24$. The difference in total accuracy rate between the individual statements and combined statements was not significant, $F(1, 56) = 0.93$, $p = .340$, $\eta_p^2 = .02$, $d = 0.18$, 95% CI [-0.20, 0.54], $BF_{10} = 0.22$. The Bayes factor analyses provided very strong evidence for the accuracy of collective statements over individual statements and over a mixture of both individual and collective statements. These findings support Hypothesis 4. Three one-way between-subjects ANOVAs were conducted to examine whether there were significant differences between the accuracy rates for truths and lie within each of the three types of statements (individual vs. collective vs. combined). Hence, these ANOVAs enabled us to check for truth or lie biases amongst the observers. The ANOVAs revealed that there were no significant differences in terms of the accuracy rates obtained for truths and lies within each of the three statement types ($F$-values ranged from 0.16 to 0.83; $p$-values ranges from .367 to .696; $BF_{10's}$ ranged from 0.20 to 0.38).

**Area Under the Curve (AUC)**

To examine how well observers were able to detect the veracity status of the pairs, areas under the curve can be estimated from the sensitivity and specificity values by applying the formula developed by Grier (1971). Hence, using the truth accuracy and lie accuracy findings above, the AUC values for the individual, collective and combined statements were .259, .867 and .481, respectively. As mentioned in Experiment 1, the cut-off criteria described by Hosmer and Lemeshow (2000) can be used as guidance for interpretation of these AUC values. Our AUC values indicate that observers were excellent at detecting veracity when using collective statements and far better at doing so with these statements than with the individual or combined statements.

**Observers' Ratings of Judgement Difficulty**

A repeated measures ANOVA was conducted to examine if there was a significant difference between Statement Type (individual vs. collective vs. combined) in terms of how difficult the observers found making their veracity judgements based on each type of statement. Statement Type did have a significant effect on observers' ratings of difficulty, $F(2, 112) = 4.35$, $p = .015$, $\eta_p^2 = .07$. Observers rated the collective statements ($M = 4.67$, $SD = 1.62$, *CI 95%* [4.24, 5.10]) as significantly more difficult for judging veracity than the individual statements ($M = 4.04$, $SD = 1.49$, *CI 95%* [3.64, 4.43]), $F(1, 56) = 4.40$, $p = .040$, $\eta_p^2 = .07$, $d = 0.40$, 95% CI [0.01, 0.75], $BF_{10} = 2.20$, and the combined statements ($M = 3.82$, $SD = 1.62$, *CI 95%* [3.40, 4.25]), $F(1, 56) = 7.41$, $p = .009$, $\eta_p^2 = .12$, $d = 0.52$, 95% CI [0.12, 0.87], $BF_{10} = 10.79$. No significant difference was found between the individual statements and combined statements in terms of difficulty in making veracity judgements ($p = .456$, $BF_{10} = 0.26$).

**What Cues do Observers Report to use?**

To explore what cues were the most relevant to observers when making their veracity judgements, 13 ANOVAs were conducted with Statement Type (individual vs. collective vs. combined) as the only within-subjects factor and each of the 13 self-reported cues as the dependent variables. The ANOVAs revealed significant effects for Statement Type for only two of the 13 cues: Verifiability and Consistency. First, Statement Type did have a significant effect on the self-reporting use of the verifiability cue, $F(2, 112) = 4.07$, $p = .020$, $\eta_p^2 = .07$. Pairwise comparisons demonstrated that observers reported using the verifiability cue significantly more frequently when judging the veracity of the collective statement (*M* =.51, *SD* = .50, *CI 95%* [.38, .64]) than when judging the veracity of the combined statements (*M* =.30, *SD* = .46, *CI 95%* [.18, .42]), $F(1, 56) = 9.14$, $p = .004$, $\eta_p^2 = .14$, $d = 0.44$, 95% CI [0.04, 0.78], $BF_{10} = 9.90$. No significant differences were found between the use of the verifiability cue for the individual statements and collective statements ($p = .088$, $BF_{10} = 0.80$) or individual statements and combined statements ($p = .350$, $BF_{10} = 0.30$). Second, a closer examination of the consistency cue revealed that no observers reported to have used it when judging the collective statements. This finding is not surprising given that observers could not use the consistency cue when judging the collective statements. Therefore, a paired samples *t*-test was conducted to examine whether there was a significant difference between the individual and combined statements based on the self-reporting use of the consistency cue. No significant difference was found between the use of the consistency cue for the individual statements (*M* =.72, *SD* = .45, *CI 95%* [.60, .84]) and combined statements (*M* =.75, *SD* = .43, *CI 95%* [.64, .87]), $p = .641$, $BF_{10} = 0.22$[ii].

**Lens Modelling: Verifiability**

The accuracy rates demonstrate that observers were significantly better at judging veracity using collective statements than when using individual statements or a combination

of individual and collective statements. We examined to what extent the observers used the verifiability approach to inform their veracity judgments of the statements. Therefore, a lens modelling style of analyses were conducted on the data to understand the influence of the number of (un)checkable details on observers' judgements of veracity. We also conducted Bayesian correlation analyses to complement the lens modelling outcomes.

In the individual statements, collective statements and combined statements, statement veracity was positively correlated with checkable details that demonstrated the pair were together (*p*-values equalled .003, < .001, and < .001 respectively; $BF_{10}$'s equalled = 24.89, 5583.16, and 6848.45 respectively) and negatively correlated with uncheckable details (*p*-values equalled < .001 for all statement types; $BF_{10}$'s equalled = 17777.69, 10547.28, and 331632.02 respectively; see Figures 1, 2 and 3). This suggests that checkable details that demonstrate the pair were together were indicative of truth-telling and uncheckable details were indicative of lying. However, the presence of verifiability cues did not influence observers' veracity judgements of the individual or combined statements: The correlations between detail presence and judgment were not significant (*p*-values for the individual statements equalled .371 ($BF_{10}$ = 0.09) for checkable details (proof pair together) and .772 ($BF_{10}$ = 0.13) for uncheckable details; *p*-values for the combined statements equalled .436 ($BF_{10}$ = 0.35) for checkable details (proof pair together) and .412 ($BF_{10}$ = 0.10) for uncheckable details; see Figures 1 and 3). Thus, when observers were rating the individual statements and combined statements they were not using the verifiability approach to judge veracity ($r$ = -.19, $p$ = .152, $BF_{10}$ = 0.07 for individual statements; $r$ = -.02, $p$ = .891, $BF_{10}$ = 0.15 for combined statements). There is clear evidence that the verifiability approach would benefit observers, but observers' judgments were made with no reference to the abundance of

(un)checkable details. Therefore, it should be of no surprise that judgments made of the individual and combined statements were inaccurate.

Conversely, what is apparent from Figure 2, is that when observers were judging the collective statements, they successfully applied the verifiability approach: Checkable details were positively correlated with veracity judgement ($r$ = .30, $p$ = .025, $BF_{10}$ = 3.82; indicative of truth-telling) and uncheckable details were negatively correlated with veracity judgement ($r$ = -.45, $p$ < .001, $BF_{10}$ = 129.32; indicative of lying). Importantly, the directions of correlations were the same on both sides of the model, an indicator of a deceptive statement led observers to judge the statement as deceptive and an indicator of a truthful statement led observers to judge the statement as truthful. This finding explains the strong correlation ($r$ = .58, $p$ < .001, $BF_{10}$ = 17568.03) between statement veracity and observers' judgment; observers were correctly using the verifiability approach when judging the collective statements.

INSERT FIGURES 1, 2 AND 3 HERE

**Lens Modelling: Consistency**

Figures 1 to 3 demonstrate that observers applied the verifiability approach when judging the collective statements only, but not when judging only the individual statements or the combination of individual and collective statements. Given that consistency was reported to have been used by observers when judging the individual statements and the combination of statements but not when judging the collective statements, we conducted further lens modelling analyses on the subjective coding of consistency data. This enabled us to examine whether observers automatically used consistency when judging multiple statements (i.e. the individual statements and the combination of statements), a cue that could not be utilised when judging only the collective statements.

In the individual statements and the combined statements, statement veracity was negatively correlated with consistency (*p*-values equalled < .001 and .008 respectively; $BF_{10}$'s equalled = 4229.63 and 4.87 respectively), which suggests that higher consistency was indicative of lying (see Figures 4 and 5). However, in the individual statements, the presence of consistency was positively correlated with observers' veracity judgments ($r$ = .33, $p$ = .013, $BF_{10}$ = 3.42), which suggests that observers associated consistent statements with truth-telling (see Figure 4). Hence, observers were utilising the consistency cue incorrectly (i.e. as a cue for truthfulness rather than deception), which explains the low accuracy rates obtained by observers when judging veracity based on the individual statements only ($r$ = -.19, $p$ = .152, $BF_{10}$ = 0.07). Conversely, the presence of consistency did not influence observers' veracity judgements of the combined statements: The correlation between consistency and judgement is not significant ($r$ = .16, $p$ = .244, $BF_{10}$ = 0.32; see Figure 5). Thus, when observers were rating the combination of individual and collective statements they were not utilising the consistency cue to judge veracity. Figure 5 suggests that if the consistency cue is used correctly then observers could benefit. However, because observers were not using consistency as a cue to determine the veracity status of the statements ($r$ = -.02, $p$ = .891, $BF_{10}$ = 0.15), the accuracy rates when judging the combination of statements achieved no better than chance.

INSERT FIGURES 4 AND 5 HERE

**Experiment 2: Discussion**

In support of Hypothesis 4, observers were able to apply the verifiability approach to better discriminate between truths and lies, but this was only the case when observers were judging one collective statement (as opposed to two individual statements or a combination of individual and collective statements). These findings were despite the fact that observers

reported that it was more difficult to judge veracity based on the collective statements than to judge veracity based on the individual statements or combined statements. In fact, the accuracy rates obtained when applying the verifiability to judge the collective statements were high (truth accuracy = 82.8%, lie accuracy = 75.0%) and amongst the highest accuracy rates obtained within lie detection research (see Bond & DePaulo, 2006 and Vrij, 2008 for reviews). These high accuracy rates demonstrate the clear potential of using the verifiability approach as well as a collective interviewing technique to detect deceit.

However, interpretation of the AUC values from the ROC curves in Experiment 1 suggest that observers should be able to classify pairs of truth-tellers and pairs of liars with high accuracy, regardless of statement type, but this was not supported by the accuracy rates or AUC values obtained in Experiment 2. Hence, the objective detection method of ROC curves outperformed human judgement when judging individual and combined statements, but was similar when judging collective statements. This is not surprising because Experiment 1 involved the systematic coding of (un)checkable details and so the ROC curves were less affected by human factors, such as bias and subjectivity. Experiment 2, however, involved non-systematic human judgements, and despite observers of Experiment 2 being taught about the verifiability approach, they still relied on other 'non-diagnostic' cues when judging the veracity of the individual and combined statements, which ultimately led to lower accuracy rates.

The fact that observers obtained a similar AUC value to that obtained by the ROC curves when judging collective statements is not surprising because observers reported to use diagnostic cues. This could mean two things. First, observers used the verifiability approach to its true potential and therefore even observers with minimal training can use the verifiability approach efficiently. Second, observers used a cue other than (un)verifiable

details to determine veracity that was highly diagnostic but not taken into account when calculating the ROC curves. We believe that the latter argument is unlikely because observers should then have performed better than they did when judging the individual and combined statements.

The lens models in Figures 1 to 5 further support this by showing that (1) the verifiability approach does work to aid the detection of deception: Observers are most accurate when they use the verifiability approach to inform their judgments of veracity; (2) observers tend to apply the verifiability approach when they only have one statement to read; and (3) when observers have access to multiple statements they make their veracity judgements using cues, such as consistency, that are not actually indicative of deceit and/or by utilising them incorrectly. Interestingly, the findings generally reflect the cues that the observers reported to have used, and as can be seen from the cues reported by observers, the main cue utilised when judging multiple statements was consistency (a cue that could not be used when judging the collective statements).

Nevertheless, whilst the lens models demonstrate that observers used the consistency cue to judge the veracity of individual statements, they used this cue incorrectly by associating consistency with truthfulness and inconsistency with deceit. Hence, observers achieved accuracy rates below chance level when rating multiple individual statements. This supports the previous research and the notion that liars actually prepare and repeat, whilst truth-tellers rely on memory and add, change and omit information over time (Granhag & Strömwall, 1999; Vredeveldt et al., 2014). Interestingly, and against what was presumed, not only did observers not apply the verifiability approach to the combined statements, but they also did not apply the consistency cue accurately. Based upon the statistical analyses conducted, it is unclear what cue(s) observers did apply to the combined statements. It is

possible that observers applied both the consistency and verifiability cues and this hampered their judgements. In fact, the accuracy rates for the combined statements suggest that the observers merely guessed (achieving accuracy rates of around 50%). When the observers had three statements (as they did in the combined condition) it is highly probable that they became confused especially because the collective statement was often a mixture of the information provided within each individual statement. Moreover, when the coders were rating the combined statements for consistency, both coders independently stated that they had difficulty when rating the consistency of the three statements (often the collective statement consisted of 50% of one individual statement and 50% of the other individual statement leading the coders to frequently rate the statements for consistency at the mid-point of 4 on the Likert scale). Therefore, although observers reported to have used consistency when judging the combined statements, it is likely that they found this difficult to apply and therefore applied both the verifiability and consistency cue or alternatively just guessed the veracity status of the pair.

Overall, the predictive accuracy of consistency is low, yet both lay people and legal professionals utilise consistency as a cue to deceit and hold the stereotypical (incorrect) belief that consistency implies truth-telling and inconsistency implies deceit (Granhag et al., 2003; Strömwall, et al., 2003; Vrij, Mann, et al, 2010). Therefore, lie detectors need to be trained to focus on alternative cues that are more predictive of deception. Such cues should be checkable and uncheckable details. Additionally, lie detectors need to be taught to avoid utilising the consistency cue when they have access to multiple statements. This will be a challenging task for real-world investigators who, as part of enforced protocol, frequently collect multiple statements during investigations. Hence, we suggest that investigators should hold off making a judgement until a collective statement has been gathered and/or

should bring in a deception detection expert to help with the investigation and the analysing of the available statements.

## General Discussion

The two experiments showed that the verifiability approach can be used to discriminate between pairs of truth-tellers and pairs of liars and that observers are able to apply this approach most reliably when judging collective statements only. We further showed that when observers judged multiple statements, they used consistency in addition or instead of verifiability. This is problematic mostly because the observers used the consistency cue incorrectly, based on the stereotypical belief that consistency implies truth and inconsistency implies deceit.

The ROC curves highlighted the potential efficiency of the verifiability approach for judging veracity regardless of statement type and therefore further research should examine whether training observers to quantify the proportions of (un)checkable details and use these measures in a standardised manner improves the ability of observers to apply the verifiability approach correctly to all statement types.

### Study Limitations

Whilst the current study demonstrates the potential of applying the verifiability approach and collective interviewing to lie detection, there are several limitations that we need to acknowledge. First, in Experiment 1, we did not manipulate the order of recall and therefore all pairs of participants first completed individual statements then completed collective statements. The reason we chose to do this sequence of recall was because we believe it to be the best order: Not only does it reflect real-life situations (e.g. at international airports or in the investigation of sham marriages), but the reverse order may affect memory

as the members of the pair will be able to discuss and jog each other's memories during the collective recall in preparation for the individual recall.

Second, the current study was a laboratory experiment involving a mock crime and a fake investigator. Whilst this is common practice in deception detection research (e.g. Nahari & Vrij, 2014; Granhag, Mac Giolla, Strömwall, & Rangmar, 2013; Vrij et al., 2009), it does lack ecological validity (a limitation of all mock crime deception studies). Consequently, although participants were not specifically informed, it is likely that the participants knew it was a mock crime, knew they were not actually being accused of a crime, and knew there was not a real investigator. Nevertheless, ethically it is not possible to conduct a real crime and in real-life it is difficult to establish ground truth, and so it is necessary to conduct laboratory studies to learn more about lie detection cues and techniques before they are applied to practice. The artificial setting of a mock crime has been studied in the context of the Concealed Information Test (CIT) whereby participants were either instructed to cheat or to cheat using their own initiative. It was found that the validity of the CIT was not restricted to instructed cheating and that the detection of cheating using the CIT was comparable across both conditions (see Geven, Ben-Shakhar, Kindt & Verschuere, 2018; Geven, Klein Selle, Ben-Shakhar, Kindt & Verschuere, 2018). These studies offer promise to the validity of deception detection laboratory studies and the applicability of such studies to the real-world. Additionally, although the current study was a laboratory study with a mock crime, it is reasonable to expect that the verifiability approach will be even stronger in real-life. This is because liars who provide false checkable details are in fact bluffing. They may bluff in experimental studies because they may think that the investigator is not likely to check the details they provide. In real-life, however, liars will probably assess the likelihood that the

investigator will check their details to be much higher; hence, this may make them less willing to bluff in real-life settings than in laboratory settings.

Third, the participants in Experiment 1 received detailed and specific instructions. The reasons for such comprehensive instructions was to make the statements of truth-tellers and liars comparable (other than veracity) and to establish ground truth. However, note that we gave instructions about the mission only, not about what truth-tellers and liars should say during the interview. Nevertheless, we cannot rule out that these detailed instructions may have influenced the recall of information. Future research could examine the impact of instructions on the ability of pairs to provide verifiable details within their statements.

Fourth, observers in Experiment 2 were naïve observers as opposed to expert investigators. Therefore, one could argue that had the observers been professionals, the findings of the current study would be different. However, previous research has shown that deception detection abilities of professionals are no better than lay people (see Vrij, 2008). Furthermore, we see no reason why the type of observer (i.e. laypersons vs. police officers) would change the accuracy differences obtained in the current study between different types of statement (e.g. collective vs. individual) nor is there any theoretical reason why professionals would be more effective at applying the verifiability approach to statements than lay people. However, future research should examine the ability of 'expert' observers (e.g. police officers) to apply the verifiability approach for identifying veracity across different types of statements, and should examine whether training observers about the limitations of the consistency cue improves their ability to successfully apply other, more diagnostics cues (e.g. checkable details) in order to more accurately judge the veracity status of statements.

Fifth, during Experiment 2 all observers were trained in the verifiability approach. This meant we had no control group to examine whether observers naturally applied this approach or to measure the degree in which the training improved performance. Although this is a limitation in experimental terms, we do not consider this to be a limitation in practical terms. A crucial prerequisite in testing the efficacy of the Verifiability Approach, and, in fact, of any lie detection method, is that users should be informed and trained in the working of the lie detection method under investigation. Furthermore, our claim that the verifiability approach improves the detection of deception is based on the finding that when participants used the verifiability approach they performed better than when they did not. It was not based on the verifiability approach training. Therefore, a control group is not necessary for our claim.

Finally, it is important to recognise that, in the current study, the verifiability approach aided the performance of observers when assessing collective statements of an alibi witness scenario only. This is just one context and therefore, until more research is conducted, it remains unclear whether the approach can be applied to other contexts. Previous research has successfully applied the verifiability approach to the gathering of information from individual suspects (Nahari et al., 2014a, 2014b).

**Conclusion**

The current study adds to the validity of the verifiability approach, demonstrating that it can be applied to alibi witness situations. Truth-telling pairs provide significantly more checkable details that demonstrate they were together (particularly during a collective statement), whereas lying pairs provide significantly more uncheckable details. When observers are taught about the verifiability approach, the collective statement facilitates their ability to more accurately distinguish between pairs of liars and pairs of truth-tellers. This

appears to be because observers do not have the option to use the consistency cue when judging only one statement, a cue which they utilise automatically especially when judging multiple individual statements.

## References

Amado, B. G., Arce, R., & Fariña, F. (2015). Undeutsch hypothesis and Criteria Based Content Analysis: A meta-analytic review. *The European Journal of Psychology Applied to Legal Context*, *7*, 3-12. DOI: 10.1016/j.ejpal.2014.11.002.

Amado, B.G., Arce, R., Fariña, F., & Vilariño, M. (2016). Criteria-Based Content Analysis (CBCA) reality criteria in adults: A meta-analytic review. *International Journal of Clinical and Health Psychology, 16*, 201-210. DOI: 10.1016/j.ijchp.2016.01.002.

Bartlett, F.C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge: Cambridge University Press.

Bond, C.F., & DePaulo, B.M. (2006). Accuracy of deception judgements. *Personality and Social Psychology Review, 10*, 214-234. DOI: 10.1207/s15327957pspr1003_2.

Burke, T. M., & Turtle, J. W. (2003). Alibi evidence in criminal investigations and trials: Psychological and legal factors. *Canadian Journal of Police and Security Services, 3*, 286-294.

Burke, T., Turtle, J.W., & Olsen, E. (2007). Alibis in criminal investigations and trials. In M.P. Toglia, J.D. Read, D.F. Ross, & R.C.L. Lindsay (Eds.), *The handbook of eyewitness psychology, Vol. 1: Memory for events* (pp. 157-174). Mahwah, NJ: Lawrence Erlbaum Associates.

Culhane, S.E., Hosch, H.M., & Kehn, A. (2008). Alibi generation: Data from U.S. Hispanics and U.S. non-Hispanic Whites. *Journal of Ethnicity in Criminal Justice, 6*, 177-199. DOI: 10.1080/15377930802243395.

Culhane, S.E., Kehn, A., Horgan, A.J., Meissner, C.A., Hosch, H.M., & Wodahl, E.J. (2013). Generation and detection of true and false alibi statements. *Psychiatry, Psychology and Law, 20*, 619-638. DOI: 10.1080/13218719.2012.729018.

Dahl, L. C., & Price, H. L. (2012). "He couldn't have done it, he was with me!": The impact of alibi witness age and relationship. *Applied Cognitive Psychology, 26*, 475-481. DOI: 10.1002/acp.2821.

Driskell, J.E., Salas, E., & Driskell, T. (2012). Social indicators of deception. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 54*, 577-588. DOI: 10.1177/0018720812446338.

Ferguson, C.J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice, 40*, 532-538. DOI: 10.1037/a0015808.

Granhag, P.A., Mac Giolla, E., Strömwall, L.A., & Rangmar, J. (2013). Counter-interrogation strategies among small cells of suspects. *Psychiatry, Psychology, and Law,* 20, 705-712. DOI: 10.1080/13218719.2012.729021.

Granhag, P.A., & Strömwall, L.A. (1999). Repeated interrogations: Stretching the deception detection paradigm. *Expert Evidence, 7*, 163-174. DOI: 10.1023/A:1008993326434.

Granhag, P.A., & Strömwall, L.A. (2000). Deception detection: Examining the consistency heuristic. In C.M. Breur, M.M. Kommer, J.F. Nijboer, & J.M. Reintjes (Eds.), *New trends in criminal investigation and evidence II* (pp. 309-321). Antwerp, Belgium: Intersentia.

Granhag, P.A., Strömwall, L.A., & Jonsson, A.C. (2003). Partners in crime: How liars in collusion betray themselves. *Journal of Applied Social Psychology, 33*, 848-868. DOI: 10.1111/j.1559-1816.2003.tb01928.x.

Grier, J.B. (1971). Nonparametric indexes for sensitivity and bias: Computing formulas. *Psychological Bulletin, 75*, 424-429. DOI: 10.1037/h0031246

Geven, L.M., Ben-Shakhar, G., Kindt, M., & Verschuere, B. (2018). Memory-based deception detection: Extending the cognitive signature of lying from instructed to self-initiated cheating. *Topics in Cognitive Science*, 1–24. DOI: 10.1111/tops.12353.

Geven, L.M., Klein Selle, N., Ben-Shakhar, G., Kindt, M., & Verschuere, B. (2018). Self-initiated versus instructed cheating in the physiological Concealed Information Test. *Biological Psychology, 138*, 146–155. DOI: 10.1016/j.biopsycho.2018.09.005.

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

Hollingshead, A. B. (1998). Retrieval processes in transactive memory systems. *Journal of Personality and Social Psychology, 74*, 659-671. DOI: 10.1037/0022-3514.74.3.659.

Home Office. (2013). *Sham marriages and civil partnerships: Background information and proposed referral and investigation scheme.* Retrieved from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/256257/Sham_Marriage_and_Civil_Partnerships.pdf.

Hosmer, D.W., & Lemeshow, S. (2000). *Applied Logistic Regression* (2nd Ed.). New York: John Wiley and Sons.

Hosch, H.M., Culhane, S.E., Jolly, K.W., Chavez, R.M., & Shaw, L.H. (2011). Effects of an alibi witness' relationship to the defendant on mock jurors' judgments. *Law and Human Behavior, 35*, 127-142. DOI: 10.1007/s10979-010-9225-5.

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. DOI: 10.7771/1932-6246.1167.

Jeffreys, H. (1961). *Theory of probability*. Oxford: UK Oxford University Press.

Karelaia, N., & Hogarth, R.M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin, 134*, 404–426. DOI:10.1037/0033-2909.134.3.404.

Lakens, D. (2016). *Power analysis for default Bayesian t-tests* [blog post]. Retrieved from: https://daniellakens.blogspot.com/2016/01/power-analysis-for-default-bayesian-t.html.

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

Loftus, E.F. (2005). Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning and Memory, 12*, 361-366. DOI: 10.1101/lm.94705.

Mossman, D. (1994). Assessing predictions of violence: Being accurate about accuracy. *Journal of Consulting and Clinical Psychology, 62*, 783-792. DOI: 10.1037//0022-006x.62.4.783.

Nahari, G., & Pazulo, 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*, 363-367. DOI: 10.1016/j.jarmac.2015.08.005.

Nahari, G., & Vrij, A. (2014). Can I borrow your alibi? The applicability of the verifiability approach to the case of an alibi witness. *Journal of Applied Research in Memory and Cognition, 3*, 89-94. DOI: 10.1016/j.jarmac.2014.04.005.

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. DOI: 10.1080/23744006.2016.1158509.

Nahari, G., Vrij, A., & Fisher, R.P. (2012). Does the truth come out in the writing? Scan as a lie detection tool. *Law and Human Behavior, 36*, 68-76. DOI: 10.1037/h0093965.

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. DOI: 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. DOI: 10.1002/acp.2974.

Potter, R., & Brewer, N. (1999). Perceptions of witness behaviour-accuracy relationships held by police, lawyers and mock-jurors. *Psychiatry, Psychology & Law, 6,* 97-103. DOI: 10.1080/13218719909524952.

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

Strömwall, L.A., Granhag, P.A., & Jonsson, A.C. (2003). Deception among pairs: "Let's say we had lunch and hope they will swallow it!" *Psychology, Crime & Law, 9,* 109-124. DOI: 10.1080/1068316031000116238.

Vernham, Z., & Vrij, A. (2015). A review of the collective interviewing approach to detecting deception in pairs. *Crime Psychology Review, 1,* 43-58. DOI: 10.1080/23744006.2015.1051756.

Vernham, Z., Vrij, A., Mann, S., Leal, S., & Hillman, J. (2014). Collective interviewing: Eliciting cues to deceit using a turn-taking approach. *Psychology, Public Policy and Law,* 20, 309-324. DOI: 10.1037/law0000015.

Vredeveldt, A., van Koppen. P.J., & Granhag, P.A. (2014). The inconsistent suspect: A systematic review of different types of consistency in truth tellers and liars. In R.Bull (Eds.), *Investigative interviewing* (pp. 183-207). New York: Springer.

Vrij, A. (2008). *Detecting lies and deceit: Pitfalls and opportunities (2nd Ed.).* Chichester: John Wiley & Sons.

Vrij, A. (2016). Baselining as a lie detection method. *Applied Cognitive Psychology, 30,* 1112-1119. DOI: 10.1002/acp.3288.

Vrij, A., Granhag, P.A., & Porter, S. (2010). Pitfalls and opportunities in nonverbal and verbal lie detection. *Psychological Science in the Public Interest, 11,* 89-121. DOI: 10.1177/1529100610390861.

Vrij, A., Jundi, S., Hope, L., Hillman, J., Gahr, E., Leal, S., Warmelink, L., Mann, S., Vernham, Z., & Granhag, P.A. (2012). Collective interviewing of suspects. *Journal of Applied Research in Memory and Cognition, 1*, 41-44. DOI: 10.1016/j.jarmac.2011.12.002.

Vrij, A., Leal, S., Granhag, P.A., Mann, S., Fisher, R.P., Hillman, J., & Sperry, K. (2009). Outsmarting the liars: The benefit of asking unanticipated questions. *Law and Human Behavior, 33,* 159-166. DOI: 10.1007/s10979-008-9143-y.

Vrij, A., Mann, S., Leal., S., & Granhag, P.A. (2010). Getting into the minds of pairs of liars and truth-tellers: An examination of their strategies. *The Open Criminology Journal, 3*, 17-22. DOI: 10.2174/1874917801003010017.

Wegner, D. M. (1987). Transactive Memory: A contemporary analysis of the group mind. In B. Mullen & G.R. Goethals (Eds.), *Theories of group behaviour* (pp. 185-208). New York: Springer-Verlag.

*Table 1: Veracity main effects obtained from the self-report data collected in the pre- and post-questioning questionnaires.*

| **Variable** | **Truth-tellers** Mean (*SD*) *95% CI* | **Liars** Mean (*SD*) *95%CI* | ***F*** | ***p*** | ***d*** (95% CI) | ***BF*$_{10}$** |
|---|---|---|---|---|---|---|
| **Pre-questioning questionnaire** | | | | | | |
| *Usefulness* | 4.92 (2.15) *4.23 – 5.60* | 6.19 (0.87) *5.88 – 6.50* | 11.40 | **.001*** ** | 0.77 (0.35, 1.10) | 10.06 |
| *Sufficiency* | 4.75 (2.09) *4.00 – 5.50* | 5.60 (1.06) *5.26 – 5.94* | 4.30 | **.042*** | 0.51 (0.11, 0.85) | 2.40 |
| *Thoroughness* | 5.33 (2.35) *4.59 – 6.08* | 5.65 (0.95) *5.32 – 6.00* | 0.61 | .437 | 0.18 (-0.19, 0.53) | 0.93 |
| *Quality* | 5.25 (2.26) *4.55 – 5.95* | 5.81 (0.85) *5.49 – 6.13* | 2.14 | .148 | 0.33 (-0.06, 0.67) | 1.39 |
| *Discussion* | 4.42 (2.19) *3.59 – 5.24* | 5.53 (1.22) *5.16 – 5.91* | 6.10 | **.016*** | 0.62 (0.22, 0.96) | $1.366^{e+9}$ |
| **Post-questioning questionnaire** | | | | | | |
| *Motivation* | 5.27 (1.70) *4.92 – 5.61* | 6.16 (0.87) *5.80 – 6.51* | 12.64 | **.001*** ** | 0.66 (0.25, 0.99) | 14.12 |
| *Confidence: Receive £10* | 6.37 (0.74) *6.09 – 6.65* | 4.93 (1.37) *4.64 – 5.22* | 50.04 | **<.001**** ** | 1.31 (0.84, 1.63) | 10990.39 |
| *Confidence: Write further statement* | 2.37 (1.38) *2.00 – 2.73* | 3.71 (1.50) *3.33 – 4.08* | 25.75 | **<.001**** ** | 0.93 (0.50, 1.25) | 20.82 |
| *Difficulty: Individual statement* | 2.58 (1.71) *2.13 – 3.03* | 3.33 (1.79) *2.87 – 3.78* | 5.30 | **.023*** | 0.43 (0.04, 0.77) | 0.56 |
| *Difficulty: Collective statement* | 2.38 (1.44) *1.97 – 2.80* | 3.55 (1.79) *3.13 – 3.98* | 15.14 | **<.001**** ** | 0.72 (0.31, 1.05) | 7.01 |
| *Truthfulness: Individual statement* | 100% (0.00) *97.68 – 102.32* | 38.45% (39.86) *36.09 – 40.81* | 1359.04 | **<.001**** ** | 2.20 (1.63, 2.52) | $2.488^{e+6}$ |
| *Truthfulness: Collective statement* | 100% (0.00) *96.23 – 103.77* | 34.14% (33.93) *30.31 – 37.97* | 590.03 | **<.001**** ** | 2.77 (2.11, 3.10) | $9.747^{e+11}$ |

*$p < .05$; **$p < .01$; ***$p < .005$; ****$p < .001$

*Table 2: Veracity X Statement interaction effect for percentage of checkable details (proof pair together).*

| | **Individual** | **Collective** | | | | |
|---|---|---|---|---|---|---|
| | Mean (*SD*) *95% CI* | Mean (*SD*) *95% CI* | ***F*** | ***p*** | ***d*** (95% CI) | ***BF*$_{10}$** |
| **Truth-tellers** | .27 (.16) *.21 – .31* | .41 (.22) *.34 – .48* | 17.67 | **<.001*** | 0.73 (0.16, 1.21) | 29.09 |
| **Liars** | .14 (.11) *.09 – .19* | .16 (.15) *.09 – .23* | 0.14 | .712 | 0.15 (-0.37, 0.66) | 0.34 |

*Table 3: Area under the ROC curves computed for each type of statement*

| **Statement** | **AUC** | **SE** | **95% CI** |
|---|---|---|---|
| All (combined) | .876* | .05 | .78, .97 |
| Individual | .789* | .06 | .67, .91 |
| Collective | .895* | .05 | .80, .99 |

*$p < .001$

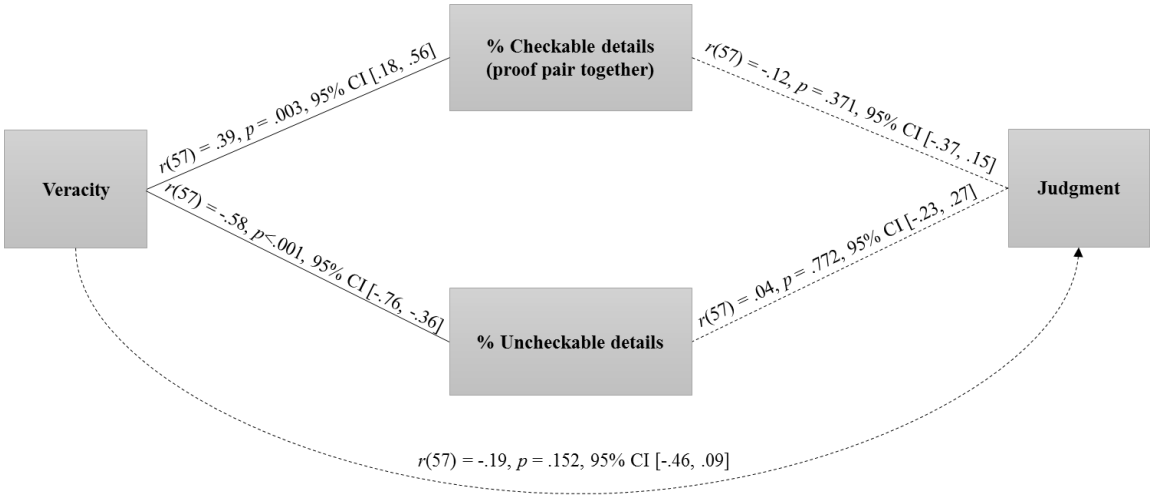


*Figure 1: Lens model for individual statements when applying the verifiability approach*

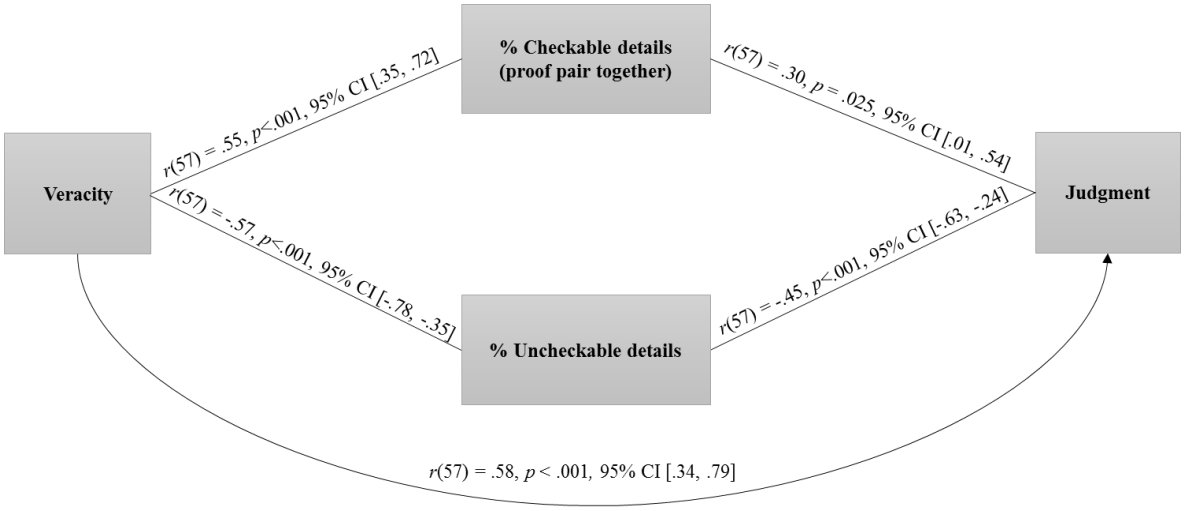


*Figure 2: Lens model for collective statements when applying the verifiability approach*

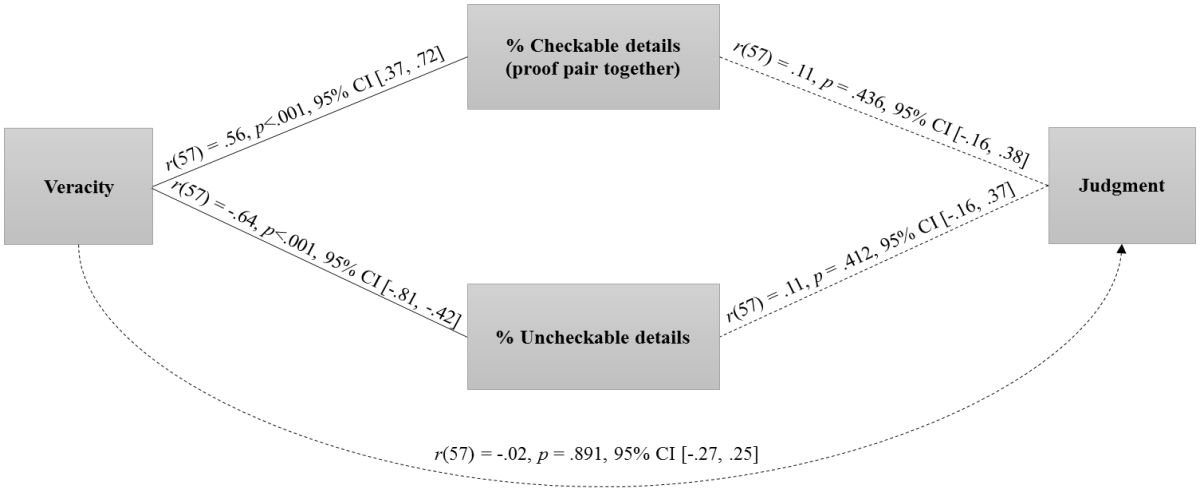


*Figure 3: Lens model for combined statements when applying the verifiability approach*

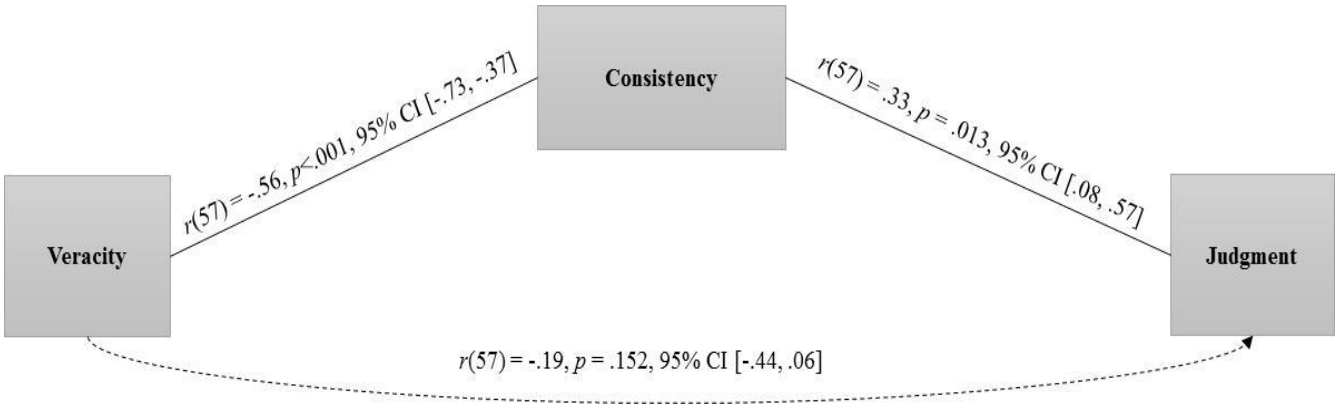


*Figure 4: Lens model for individual statements when applying consistency*

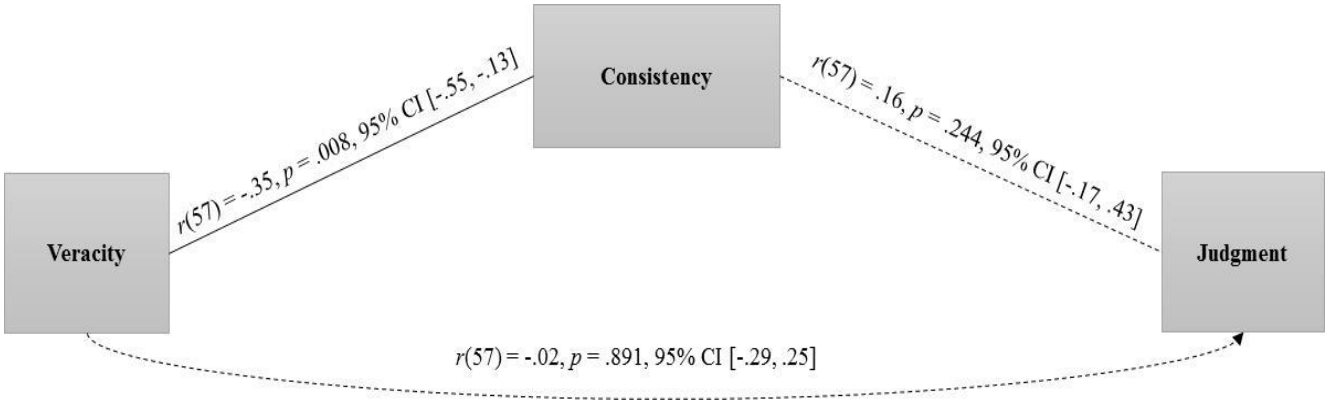


*Figure 5: Lens model for combined statements when applying consistency*

[i] Checkable details (other) was not included as a variable in this second experiment because, in the first experiment, we did not predict, or find, any differences between truth-tellers and liars regarding this variable.

[ii] The 11 ANOVA's that were conducted on the remaining 11 cues that were mentioned by observers when making their veracity judgements revealed no significant findings: Statement type did not have an effect on the self-reporting of number of details, pronouns, plausibility, unnecessary details, feelings/opinions, equality, interactions, statement length, overcomplicated information, specificity of information, or memory recall.