Applied Cognitive Psychology

WILEY

RESEARCH ARTICLE OPEN ACCESS

Noise in the Verifiability Approach to Lie Detection

Germán Gálvez-García[1,2] | Patricio Mena-Chamorro[2] | Alejandro Moliné[3] | Tomás Espinoza-Palavicino[2] | Oscar Iborra[3] | Emilio Gómez-Milán[3]

[1]Departamento de Psicología Básica, Psicobiología y Metodología de las Ciencias del Comportamiento, Facultad de Psicología, Universidad de Salamanca, Salamanca, Spain | [2]Departamento de Psicología, Universidad de La Frontera, Temuco, Chile | [3]Centro de Investigación Mente, Cerebro y Comportamiento, Universidad de Granada, Granada, Spain

Correspondence: Germán Gálvez-García (germangalvezgarcia@usal.es)

Received: 27 December 2024 | **Revised:** 3 June 2025 | **Accepted:** 16 June 2025

Funding: This work was supported by the Fondo Nacional de Desarrollo Científico y Tecnológico, 1230431 and Agencia Nacional de Investigación y Desarrollo, 21210298, 21231269.

Keywords: algorithmic classification | anti-noise protocol | consensus judgments | deception detection | verifiability approach

ABSTRACT

This study examined the effectiveness of Verifiability Approach (VA) in deception detection. Additionally, we examined potential sources of judgmental noise, including subjective variability in detail counts and inconsistencies in truthfulness assessments. Two experiments were conducted. In the first, untrained participants counted verifiable and unverifiable details in alibi statements without assessing truthfulness, revealing variability in detail counts as a significant noise source. In the second, participants assessed truthfulness using four methods: personal judgment, verifiable detail-assisted judgment, algorithmic classification, and a new consensus-based judgment protocol. Algorithmic classification improved accuracy (61.00%) compared to personal and assisted judgments but resulted in a high false alarm rate (59%). The consensus-based judgment or anti-noise protocol reduced false alarms (30%) and showed promise for enhancing reliability. While the VA enables detection of deception above chance level, its effectiveness is limited by subjective judgment biases. Providing judge training and applying an anti-noise protocol may improve its reliability.

1 | Introduction

Deception is a pervasive feature of human communication (Cole 2001; DePaulo et al. 1996; Jensen et al. 2004; Serota et al. 2010). This concern has prompted the development of deception detection methodologies to ensure judicious appraisal (Masip 2017; Verschuere et al. 2023). From a behavioral approach, the identification of nonverbal cues (e.g., facial micro-expressions) and verbal content analysis of individuals has been proposed as detection methods (Bogaard et al. 2016a, 2016b, 2019; Denault 2020; Denault et al. 2020; Granhag and Hartwig 2015; Köhnken et al. 2015; Sporer and Schwandt 2007; Vrij 2019; Vrij et al. 2018, 2016, 2022). However, the accuracy of these methods is often limited, and the evaluators' cognitive biases may compromise the objectivity of judgments (Hartwig and Bond Jr 2011). Meta-analyses have demonstrated that individuals' ability to detect lies generally does not exceed random chance (Bond and DePaulo 2008; Hartwig and Bond Jr. 2014; Hauch et al. 2016; Palena et al. 2021). Nahari et al. (2014a, 2014b) proposed the Verifiability Approach (VA) to overcome this challenge.

The VA evaluates statements based on three dependent variables: (1) the number of verifiable details, which are specific and confirmable facts (e.g., 'I was at the café with my sister at 3 PM'); (2) the number of unverifiable details, which lack external confirmation (e.g., 'I was thinking about my plans for the weekend'); and (3) the ratio of verifiable to unverifiable details, which provides a relative measure of the reliability of the statement. These variables serve as key indicators in

distinguishing truth from deceptive accounts. The VA is based on the premise that people who tell the truth are more likely to provide specific and verifiable details about the events they describe. Conversely, liars face a dilemma: providing too few details may raise suspicion, but offering too many details may increase their vulnerability to detection, as inconsistencies could arise during scrutiny (Nahari 2018; Nahari et al. 2014a, 2014b; Strömwall et al. 2006; Vrij et al. 2010). Thus, lies often include details that are difficult or impossible to verify (Verschuere, Bogaard, and Meijer 2021). During interrogation, suspects provide verifiable and unverifiable details voluntarily and in response to questions. Investigators classify these details accordingly after taking the statement or conducting further investigations. The distinction between detail is clarified by Vrij and Nahari (2019), defining four types of verifiable details: (a) activities that were carried out with identifiable people whom the interviewer can consult (e.g., "I went to a restaurant with my sister"); (b) activities witnessed by witnesses who can be identified and whom the researcher can question (e.g., "When I was leaving the house, my neighbor was coming in and saw me leaving"); (c) actions that the interviewee believes were recorded by closed-circuit television (CCTV) (e.g., "I took money out of the ATM at 5 p.m. and I had a camera"); (d) actions that could have been recorded in some other way (e.g., "I called my friend at 1:45 p.m."). On the other hand, unverifiable details lack these characteristics and represent a weak index of lying since they may also depend on the declarant's motivation level.

This general framework has been empirically tested across numerous studies. One of the most comprehensive evaluations of the VA to date is the meta-analysis by Palena et al. (2021), which showed that truth tellers reported significantly more verifiable details than liars ($g = 0.42$). In contrast, unverifiable details did not reliably differentiate between the two groups ($g = -0.25$). They also found that the diagnostic utility of the VA increased under specific boundary conditions, especially when the procedure included an information protocol. In this protocol, participants were explicitly informed about the rationale behind the VA and told that their statements would be checked. Under these conditions, the effect size increased substantially when written statements focused on event-related content ($g = 0.80$). Building on these findings, the present study follows all methodological recommendations provided by Palena et al. (2021) to maximize the effectiveness of the VA. However, our main focus lies not on the behavior of the interviewees but on the variability introduced by human raters when applying the VA, which remains a relatively unexplored yet important source of noise in its application.

Regarding empirical studies using the VA, Verschuere, Bogaard, and Meijer (2021) found that the VA discriminated more accurately about the veracity of different alibi statements than any other cue, which participants considered useful; the results of Verschuere, Bogaard, and Meijer (2021) were corroborated by Verschuere, Schutte, et al. (2021) in a second study where students were asked to tell what they had done in the last 15 min on campus. The truth-tellers took part in regular activities on campus (i.e., calling a friend, drinking coffee, etc.). At the same time, the liars had committed a simulated crime (i.e., they had stolen a test but falsely claimed they were innocent and had been on campus doing activities). In this study, those who evaluated a single cue (i.e., data verifiability or richness of detail) obtained surprisingly high judgment accuracy. This suggests that focusing on the most reliable cue yields better accuracy than considering multiple cues. More recently, Verschuere et al. (2023) asked their participants: "How verifiable is this statement on a scale ranging from totally unverifiable (–100) to totally verifiable (+100)?". They obtained an accuracy of 70%. In other studies, it was found that truth-tellers provided much more verifiable details than liars (Bogaard et al. 2020; Jupe et al. 2017; Harvey et al. 2017; Deeb et al. 2021; Boskovic et al. 2017; Vernham et al. 2020).

Although the VA provides a structured framework, it is susceptible to *noise*—inconsistent variations in judgment caused by subjective biases, differing interpretations, or external influences (Kahneman et al. 2021). For instance, two judges might classify various verifiable details or judge the exact count differently based on their individual thresholds for truth or deception. To mitigate such inconsistencies, Kahneman et al. (2021) propose an *anti-noise protocol*. This protocol involves structuring decision-making into independent assessments, using a standard reference frame, organizing information sequentially, and averaging independent judgments to enhance consistency and objectivity. It should be noted that the free recall conditions must be bound in time (N minutes for the statement) and similar for all the declarants. Finally, the average number of verifiable details provided by a set of independent judges on a story should be compared with an algorithmic classification rule (i.e., a criterion of minimum number of verifiable details to be considered true), based on a database of true stories of the same type as the evaluated story and obtained under similar conditions. However, despite these clear guidelines, there remains a gap between theoretical recommendations and actual implementation in lie detection research.

Despite the potential benefits of the VA, many recommended anti-noise procedures are rarely applied in practice. Palena et al. (2021) identified several moderators that increase the effectiveness of the VA, such as the use of written, event-related statements and informing participants in advance that their statements would be checked. However, most previous studies have relied on truthfulness judgments made by a single judge or by a small group of judges who reached consensus. This practice may give the impression of objectivity but can mask individual differences in judgment. As Kahneman et al. (2021) argues, agreement reached through discussion often creates an illusion of reliability. This issue is visible in Verschuere, Schutte, et al. (2021) and in most of the 14 studies included in the meta-analysis by Palena et al. (2021). Methodological decisions like these can obscure the variability in how verifiability is interpreted and assessed. Palena et al. (2021) also reported that effect sizes for verifiability cues were lower when assessors were not trained or independent verification procedures were not used. This suggests that the effectiveness of the VA can be substantially compromised when these protocols are not rigorously implemented. In the present study, we aim to address these limitations directly by applying an anti-noise protocol designed to reduce rater-based variability. Our focus is not only on the verbal behavior of the participants, but also on the decision-making

process of the assessors and the potential noise arising from unstandardized verifiability judgments.

As Kahneman et al. (2021) emphasize, different types of noise in evaluative decisions, such as judgments of truth or deception, are randomly distributed. These include occasion noise, referring to temporary factors like a rater's mood. System noise stems from stable characteristics such as personality traits, empathy, or beliefs about one's lie-detection ability. Pattern noise arises from interactions between a specific judge and a specific case, including biases. Random noise can only be reduced through the implementation of decision hygiene. One of the most effective hygiene strategies is averaging the judgments of several independent raters or using an algorithmic approach. However, the first solution introduces a practical trade-off: to preserve the low-cost and efficient nature of the VA as a lie detection tool, the number of raters must be kept as small as possible. According to Kahneman et al. (2021) the starting point must always be noise auditing. This also applies to the VA. Identifying and analyzing sources of variability in judgments of truth and deception is essential for improving reliability. Among these sources, the first one to be ruled out is incompetence. This may result from inexperience in novices or overconfidence in experienced raters who have practiced without feedback. This is especially relevant in real-world contexts, where feedback on the accuracy of truth judgments is often unavailable and confidence-related bias and noise can become dominant. Our first step in this direction is to investigate judgmental incompetence in novices. We focus on the judgment process itself, as evaluating how judgments are formed is considered one of the most robust anti-noise strategies, both for verifiable and unverifiable content, as suggested by Kahneman et al. (2021).

Accordingly, the present study systematically investigates the main sources of noise associated with VA and their impact on lie detection. Specifically, we pursue two objectives: (a) to test whether the VA improves lie detection performance in novice raters (Experiments 1 and 2), and (b) to evaluate how different noise sources in the application of the VA affect the accuracy of deception judgments (Experiments 1 and 2). These objectives were addressed through two experiments in which participants assessed the truthfulness of alibi statements describing recent campus activities of college students.

2 | Experiment 1: The Role of Noise in the Verifiability Approach for Counting Verifiable Details

In Experiment 1, participants counted verifiable and unverifiable details of alibi statements without judgment on their truthfulness or deception. Since participants did not judge the statements, a cut-off point was established based on the anti-noise protocol of Kahneman et al. (2021). Specifically, the mean number of verifiable details of statements across N independent judges was used to classify statements as truthful. Alibi statements with a number of verifiable details higher than this mean were considered truthful. The first objective of this experiment was to evaluate whether the VA improves deception detection performance compared to chance in untrained participants. Although the judges had no formal training, they received theoretical information about the VA, including definitions of verifiable and unverifiable details, and completed two example trials with feedback to practice the detail-counting task. The second objective was to examine the variability and consistency in the count of verifiable and unverifiable details since it is a potential noise source in the evaluations of untrained judges.

2.1 | Method

2.1.1 | Participants

Eighteen students from the University of Granada (50% women; mean age = 24.9 years, SD = 5.2; age range = 18–32 years) were recruited for this experiment through convenience sampling. An a priori power analysis was conducted using G*Power (version 3.1; Faul et al. 2009) to determine the sample size, assuming a medium effect size ($f^2 = 0.25$), a moderate correlation between measures ($r = 0.4$), 95% statistical power ($1 - \beta = 0.95$), and a significance level of 5% ($\alpha = 0.05$) (*F*-test family, repeated-measures ANOVA). This analysis indicated a minimum of 16 participants; however, 18 healthy adults were recruited to enhance the robustness of statistical analyses. This slight oversampling enhanced statistical stability and reduced Type I and II error risks, particularly in analyzing additional variables. Participants were informed about the general aim of the study. They gave informed consent but were not aware of the specific role of the VA until after completing the experiment. Additionally, they did not receive any monetary compensation.

2.1.2 | Materials

A total of 20 alibi statements (10 truthful and 10 deceptive) were used in this study. Statements were manually anonymised, with real names or identifiable information replaced by fictitious alternatives. These statements were selected from a larger set generated under controlled experimental conditions in our previous research. In those studies, participants provided a 2-min statement about their activities during the previous 15 min on the campus of the Faculty of Psychology via telephone call. Participants either described their activities truthfully or fabricated a story. In both the true and false statement conditions, a research assistant served as the designated witness, ensuring content accuracy or falsity. Following the study by Verschuere, Schutte, et al. (2021), the deceptive participants were given a tutorial on how to answer the question "*What have you done in the last 15 min on campus?*" in the office of an assistant to the principal investigator. In the call, these participants were instructed to lie about activities they had not actually done.

All statements were transcribed, and the truthful and deceptive texts were compared in terms of coherence and length. Coherence was rated by two independent judges on a 3-point Likert scale (1 = incoherent, 2 = partially incoherent or containing incoherent details, 3 = coherent). Length was measured as the average number of words per minute. No significant differences were found between truthful and deceptive statements in either variable. Mean coherence scores were 2.8 (SD = 0.3), and

the mean word rate was 115 words per minute (SD = 14). The average statement duration was 120 s (SD = 21).

2.1.3 | Procedure and Settings

At the time of recruitment, participants were informed that they would participate in a deception detection experiment. Upon arrival at the laboratory, participants were given an overview of the study's aim. They were provided with a brief overview of the VA and informed that their statements would be analyzed using this method. However, they were not made aware of its specific aspects (e.g., noise sources) until after completing the experiment. Participants were incentivized to lie by offering course credit, which would be doubled if their deceptive statement was not detected during later evaluation. Participants were asked to count the verifiable and unverifiable details in 20 alibi statements after two practice trials (one truthful and one deceptive), which included feedback, including the group average and the actual number of verifiable and unverifiable details for each statement. This was done to familiarize participants with the complexity of the task and the distinction between detail types. Counting took place in an isolated room in the laboratory, and participants were not asked to judge the truthfulness of the statements. The presentation of truthful and deceptive alibi statements was randomized. After completing the count of details for each of the 20 test statements, participants did not receive feedback regarding the effectiveness of their judgments based on the count of verifiable and unverifiable details. The experiment was completed in a single 45-min session.

2.1.4 | Data Analysis

All data analyses were performed in RStudio (version 2024.09.0; RStudio Team 2020) using several data processing and modeling packages. The accuracy and false alarm rates were calculated to evaluate the lie detection performance of the VA. Chance overcoming was considered with a minimum of 60% accuracy and a maximum of 40% false alarms (Gálvez-García et al. 2021). For each of these indices, the number of verifiable details in each declaration was calculated by comparing the number of verifiable details in each declaration with a cut-off point based on the average number of verifiable details in statements and judges. Following Kahneman et al. (2021), the arithmetic mean is considered the most reliable estimator for minimizing noise, particularly when judgments are made independently, minimizing overall error, and practically: when judgments are independent, the mean reduces noise, whereas it may increase it in deliberative contexts. In our experiments, raters made independent judgments. Statements with a number of verifiable details higher than this cut-off point were classified as "truthful" and statements with a lower number were classified as "deceptive" Accuracy was defined as the percentage of instances in which the average number of verifiable details correctly identified deceptive statements as deception. False alarms were defined as the percentage of occasions when the average verifiable detail incorrectly identified truthful statements as deceptive.

In order to determine whether the percentage accuracy obtained was statistically different from chance (50%), a binomial test was performed on the percentage accuracy of the average of verifiable details. Furthermore, linear mixed-effects models (LME) were estimated to assess the variability in the count of verifiable and unverifiable details. LME models were computed with the *lme4* (Bates et al. 2015) package. The models included the factors 'statement veracity' (i.e., truthful and deceptive) as fixed effects and 'participants' as a random effect. Model selection was performed using a comparison approach (McElreath 2016). Likelihood ratio tests and Akaike's information criterion (AIC) (i.e., lower AIC values indicate better model fit) were used to determine whether including fixed effects or interactions improved the fit. In addition, the relative likelihood between models was calculated using the exponential of the difference in AIC [AICRL = exp.(ΔAIC/2)] to assess the parsimony of each model (Espinoza-Palavicino et al. 2023; Gálvez-García et al. 2024). The complete model contrasted with the null model, which included only a random effect (e.g., Model A: Number of verifiable details ~ Statement veracity + 1|Participants versus Model B: Number of verifiable details ~1|Participants).

Planned comparisons were estimated with the *emmeans* package (Lenth 2023), and the Tukey method was used to adjust *p*-values. In addition, marginal and conditional R^2 coefficients were calculated with the *piecewise SEM* package (Lefcheck 2016) to account for the variance explained by fixed effects (i.e., marginal R^2) and joint fixed and random effects (i.e., conditional R^2). These values allowed the interpretation of each model's contribution of fixed effects.

Finally, different intraclass correlation coefficients (ICC) were calculated to assess consistency in the count of verifiable details. To analyze the consistency of an individual judge across all statements (i.e., judge one across all statements), the *Fixed Effects Model for Raters* (ICC3) was used. This model assumes that judges are fixed and that their evaluations do not generalize to other judges. To assess the average consistency among all judges across all statements (i.e., all judges across all statements), the *Random Effects Model for Average Raters* (ICC2k) was used. This model assumes that judges are interchangeable and that their ratings could be generalized to other judges. Additionally, Pearson's correlation coefficients were calculated to evaluate the pairwise association in the count of verifiable details between judges. Consensus was defined as majority agreement: a classification was considered consensual when at least two of the three judges agreed in their veracity judgment. This analysis provides additional insight into the degree of agreement between individual judges and complements the intraclass correlation coefficients (ICC), which assess overall consistency across judges. Together, these analyses allowed us to assess both classification accuracy and inter-rater consistency, as well as the contribution of statement veracity to rating variability.

2.2 | Results

2.2.1 | Accuracy and False Alarm Rates

Based on a total of 180 classification trials (18 participants × 10 statements), the average number of verifiable details in statements ($M = 3.06$, SD = 2.43) was used as a cut-off point to calculate accuracy and false alarm rates. Using this classification

algorithm, the judges achieved an accuracy rate of 64.44% and a false alarm rate of 57.78%. Furthermore, the binomial test showed that the observed number of correct classifications (i.e., 116 successes out of 180 trials) was significantly higher than what would be expected by chance, indicating that the accuracy rate of 64.44% was significantly different from random guessing ($p < 0.001$). This suggests that the verifiable details classification algorithm performed better than random guessing in detecting truthful and deceptive statements.

2.2.2 | Inter-Judge Variability Between Verifiable and Unverifiable Details

The LME model for verifiable details (see Figure 1A) showed that the main effect of statement veracity was not statistically significant ($\chi^2_{(1)} = 2.738$, $p = 0.097$, $AIC_{RL} = 0.691$, $R^2_m = 0.005$, $R^2_c = 0.382$). This non-significant result indicates that the inclusion of statement veracity did not improve model fit over the null model, as also suggested by the AICRL and marginal R^2. By contrast, the LME model for unverifiable details (see Figure 1B) revealed a significant main effect of statement veracity ($\chi^2_{(1)} = 7.733$, $p = 0.005$, $AIC_{RL} = 0.056$, $R^2_m = 0.017$, $R^2_c = 0.235$), indicating that the inclusion of statement veracity as a fixed effect improved model fit over the null model. Planned comparisons showed that the number of unverifiable details was significantly higher for truthful statements ($M = 3.006$, $SD = 2.717$) compared to deceptive statements ($M = 2.394$, $SD = 1.930$, $t_{(343)} = 2.793$, $p = 0.005$).

2.2.3 | Judge Consistency in Verifiable Detail Counts

For individual judges, the fixed effects model for raters (ICC3) showed low and statistically significant consistency ($ICC3 = 0.21$, $p < 0.001$). This suggests the same judge's judgments were inconsistent in counting verifiable details across truthful and deceptive statements. On the other hand, for all judges, the random effects model for average raters (ICC2k) showed moderate and statistically significant consistency ($ICC2k = 0.73$, $p < 0.001$). This indicates a high level of consistency among judges when evaluating both truthful and deceptive statements. For instance, in one deceptive statement, the number of verifiable details counted by judges ranged from 0 to 12 ($M = 3.11$, $SD = 3.00$). This wide range exemplifies the low intra-rater consistency observed, as reflected in the ICC3 value.

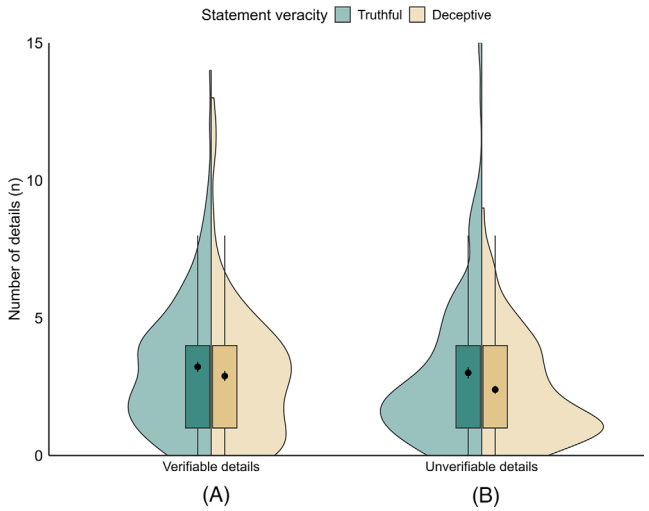

FIGURE 1 | Experiment 1. (A) Average number of verifiable details for Statement veracity; (B) Average number of unverifiable details for Statement veracity. Error bars represent standard errors of the mean. *$p < 0.05$, **$p < 0.010$, ***$p < 0.001$, ns (not significant).

The correlation matrix showed that the associations between judges were mostly low to moderate ($r_{mean} = 0.228$), with a few high correlations ($r = -0.37$ to 0.97). Furthermore, only a few of these associations were statistically significant (see Figure 2). These findings indicated high variability in the judges' counting of verifiable details.

2.3 | Discussion

The first objective of experiment 1 was to assess whether the VA to deception detection improves performance compared to chance. The findings showed that the VA achieved an accuracy rate of 64.44%, significantly higher than that expected by chance (50%). This provides preliminary evidence that the VA can improve performance over random guessing, even in untrained raters. However, a high false alarm rate (57.78%) was also observed. Although the verifiable detail classification algorithm effectively identifies deceptive statements, it also misclassifies many truthful statements as deceptive. This pattern could be partly explained by the lack of training and feedback on counting verifiable details among the participants in this study. The absence of clear criteria to differentiate between verifiable and unverifiable details could lead judges to overestimate the number of verifiable details, especially in deceptive statements.

The second objective of this experiment was to assess the variability and consistency in counting verifiable and unverifiable details by untrained judges. The results indicated that judges tend to count significantly more unverifiable details in truthful statements than in deceptive statements, with no differences observed in the count of verifiable details between the two types of alibi statements. This supports the idea of focusing the analysis of the VA on verifiable details rather than using the ratio of verifiable to unverifiable details. The number of verifiable details did not significantly differ between truthful and deceptive statements, indicating that this metric remains relatively constant across conditions. In contrast, the ratio between verifiable and unverifiable details could introduce more noise and lead to less accurate conclusions. This is because truthful statements often include more unverifiable details, and identifying such details tends to vary across judges. These inconsistencies in judging unverifiable content can distort the ratio metric. Therefore, focusing only on verifiable details, defined here as the mean number of verifiable items identified by all judges for a given statement, offers a more stable and reliable measure. Similarly, the 'judges' average' refers to the aggregated binary classification of truthfulness based on multiple raters' independent judgments. These findings are consistent with the study by Vrij and Nahari (2019), who showed that unverifiable details might depend on the level of motivation of those who lie, hypothesizing that only liars with a high level of motivation for not being caught would show

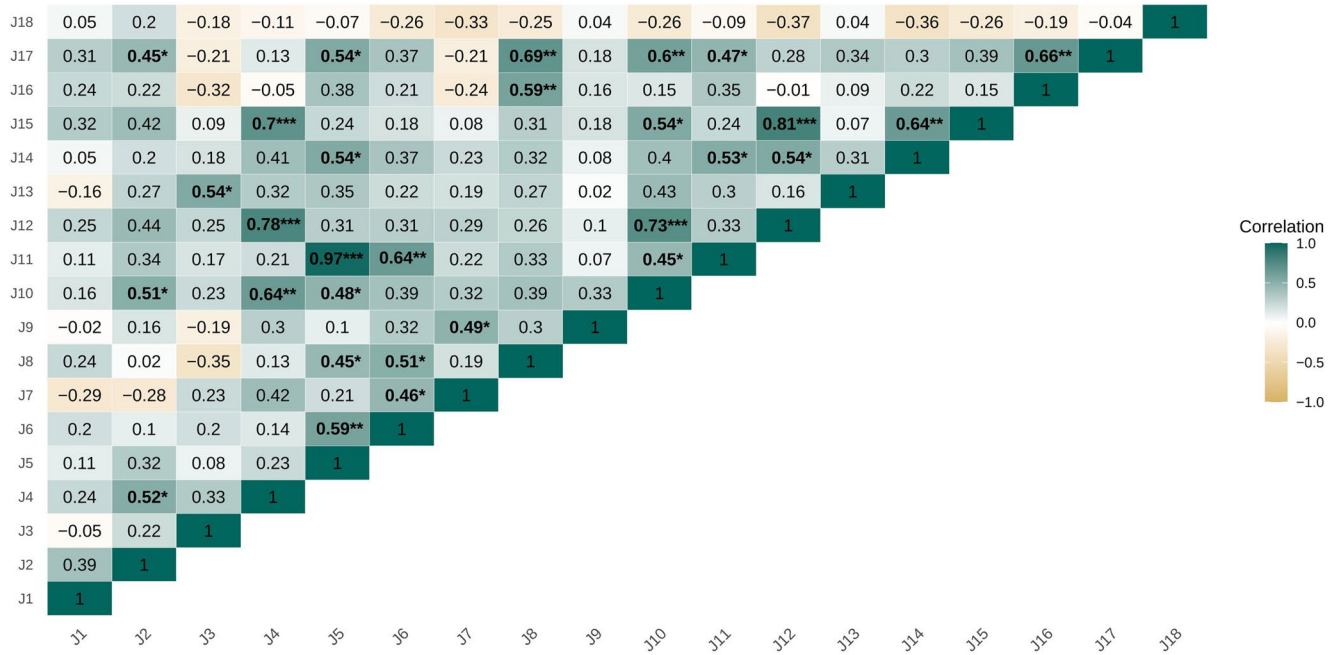

FIGURE 2 | Experiment 1. Pearson's correlation matrix between judges' verifiable detail counts. Bold values indicate statistical significance ($p < 0.05$*, $p < 0.01$**, $p < 0.001$***).

unverifiable details in their accounts. This suggests that unverifiable details represent a misleading index of lying. In addition, there was high variability in the count of verifiable details when the same judge evaluated different statements, while consistency between judges was moderate. These findings offer an additional explanation for the high false alarm rate observed since the high variability in the count of verifiable details could indicate that some judges were less rigorous in their individual assessments.

In this sense, using a classification algorithm based on the mean of verifiable details of truthful statements allows reducing the noise associated with the count of verifiable details in alibi statements, similar to the anti-noise protocol of Kahneman et al. (2021). In Experiment 2, we will specifically isolate and evaluate how subjective human judgment contributes to overall noise in veracity assessments within the VA.

3 | Experiment 2: The Role of Noise in the Verifiability Approach for Judgment

In Experiment 2, participants counted verifiable details of alibi statements and provided judgments of truthfulness or deception. While Experiment 1 addressed the first source of noise in the VA (i.e., variability in counting verifiable details), Experiment 2 focused on a second source: the judgment of truthfulness or deception. Four methods were evaluated and compared: (a) the truthfulness judgment conducted by the judges according to their own count of verifiable details (method A: personal judgment), (b) the judgment conducted by each of the judges after having known the average value of verifiable details corresponding to each statement (method B: verifiable detail-assisted judgment), (c) the judgment generated by the verifiable details classification algorithm (method C: algorithmic classification; it uses as the cut-off point the overall average of verifiable details across all the texts), and (d) judgment based on a new consensus-based classification protocol using the average of judges' responses, developed to assess the truthfulness of alibi statements (method D: consensus-based judgment protocol).

This consensus-based anti-noise protocol (method D: consensus-based judgment protocol) classifies a statement's veracity according to the majority of individual judgments. A count was made of how many judges classified each statement as 'true' or 'false' in method B: verifiable detail-assisted judgment. Then, a majority rule is applied; if more than half of the judges consider a statement to be "true," the rule classifies the statement as true; otherwise, it classifies it as "false." This approach aims to reduce the impact of possible individual biases and reflect a collective judgment derived from inter-rater agreement. In other words, method D: consensus-based judgment protocol, implements a two-tier anti-noise protocol: one average for the verifiable detail count and another for the truthfulness judgment. The first objective was determining which method most effectively reduces judgment-related noise and enhances deception detection accuracy. The second objective was to analyze the variability in truthfulness or deception judgments made by untrained judges.

3.1 | Method

3.1.1 | Participants

Twenty new students from the University of Granada (75% women, mean age = 23.4 years, SD = 6.2; age range = 18–32 years) were recruited through convenience sampling. The sample size was determined using the same a priori power analyses described in Experiment 1 (see Section 2.1.1). Participants gave informed consent after receiving a full explanation of the study's

aims and procedure. As in Experiment 1, no monetary compensation was offered.

3.1.2 | Materials

A total of 20 new alibi statements (10 truthful and 10 deceptive) were used in Experiment 2. These statements were randomly selected from the same set of statements described in Experiment 1 (see Section 2.1.2).

3.1.3 | Procedure

The procedure employed in this experiment was similar to that of Experiment 1 (see Section 2.1.3) in counting verifiable details. However, in this experiment, participants both counted verifiable details and made judgments of truthfulness. At the time of recruitment, participants were informed that they would participate in a deception detection experiment. Upon arrival at the laboratory, participants were given an overview of the study's aim and a brief overview of the VA. However, they were not made aware of its specific aspects (e.g., noise sources) until after completing the experiment. As in Experiment 1, participants completed two practice trials (one truthful and one deceptive) to familiarize themselves with the detail-counting task before starting. Then, participants counted the verifiable details in 20 alibi statements, presented randomly in a controlled environment in an isolated room. After counting, participants assessed the truthfulness of the same statements at two points in time: first, based solely on their own counts (method A: personal judgment), and second, after knowing the average count (method B: verifiable detail-assisted judgment). Importantly, participants did not receive feedback on the accuracy of their judgments, and the entire experiment was completed in a single session lasting approximately 55 min.

3.1.4 | Data Analysis

The data analyses used in this experiment were similar to Experiment 1 (see Section 2.1.4) regarding the accuracy and false alarm calculations, estimation of LME models, and consistency in truth and deception judgments. All statistical analyses were performed in the RStudio program (version 2024.09.0; RStudio Team 2020). Accuracy and false alarm rates were calculated for judgments based on participants' own counting (method A: personal judgment), judgments provided after knowing the average verifiable detail value corresponding to each statement (method B: verifiable detail-assisted judgment), judgments provided by the verifiable detail classification algorithm (method C: algorithmic classification), and method D: consensus-based judgment protocol. It should be noted that method D could not be included in the statistical comparisons due to its data structure. This consensus-based classification yields a single binary decision per item, with no variability between individual judgments, rendering it incompatible with the statistical models applied to individual or algorithmic classifications. Consensus was defined as a simple majority (i.e., at least two out of three raters agreeing on the classification). All raters worked independently and were blind to each other's decisions.

As in experiment 1, the binomial test was used to determine whether the accuracy percentage obtained from the different judgments was statistically different from chance (50%). Also, LME models were estimated to compare the accuracy and false alarms of the different judgments. The models included the factors "method" as fixed effects and "participants" as a random effect. The likelihood ratio test, AIC and the relative probability of AIC were used to choose the best model. The complete model was contrasted against the null model (e.g., Model A: Percent Accuracy ~ Method +1|Participants versus Model B: Percent Accuracy ~1|Participants). Planned comparisons were performed using the Tukey method. In addition, to calculate the variance explained by the fixed and random effects, the marginal and conditional R^2 coefficients were estimated, respectively. An LME model was also estimated to assess variability in the count of verifiable details. The model included 'statement veracity' (i.e., truthfulness and deceptiveness) as a fixed effect and 'participants' as a random effect. The same criteria were used to choose the best model, test the model, perform planned comparisons, and calculate the explained variance of the fixed and random effects. Finally, ICC3 and ICC2K were calculated to assess individual judge consistency and averaged across all judges in counting verifiable details, conducting personal judgments (method A), and making judgments assisted by verifiable details (method B), respectively. Additionally, Pearson correlation coefficients were calculated to evaluate, on the one hand, the association between the judgments made by all the judges and, on the other hand, the associations in the count of verifiable details between them. This analysis provides an additional perspective on the level of agreement between judges, complementing the information obtained through the ICC coefficients.

3.2 | Results

3.2.1 | Accuracy and False Alarm Rates by Methods

Performing the judgment by their own count of verifiable details (method A: personal judgment), achieved an accuracy rate of 51.50% and a false alarm rate of 52.50%. The binomial test showed that the number of correct trials (i.e., 103 successes out of 200 trials) was not significantly higher than expected by chance. This indicates that the accuracy rate of 51.50% is not significantly different from random guessing ($p = 0.723$).

As for the judgment performed after knowing the average number of verifiable details (method B: verifiable detail-assisted judgment), the judges achieved an accuracy rate of 53.50% and a false alarm rate of 44.50%. The binomial test showed that the number of correct judgments (i.e., 107 successes out of 200 trials) was not significantly higher than expected by chance. This suggests that the accuracy rate of 53.50% is not significantly different from the random guess ($p = 0.358$). As for the verifiable detail classification algorithm (method C: algorithmic classification), the average number of verifiable details in truthful statements (cutoff point: $M = 2.33$, SD = 1.66) was used to calculate accuracy and false alarm rates. The algorithm achieved an accuracy rate of 61.00% and a false alarm rate of 59.00%. The binomial test showed that the number of correct trials (i.e., 122 successes out of 200 trials) was significantly higher than expected by chance. This suggests that the 61.00% accuracy rate significantly differed

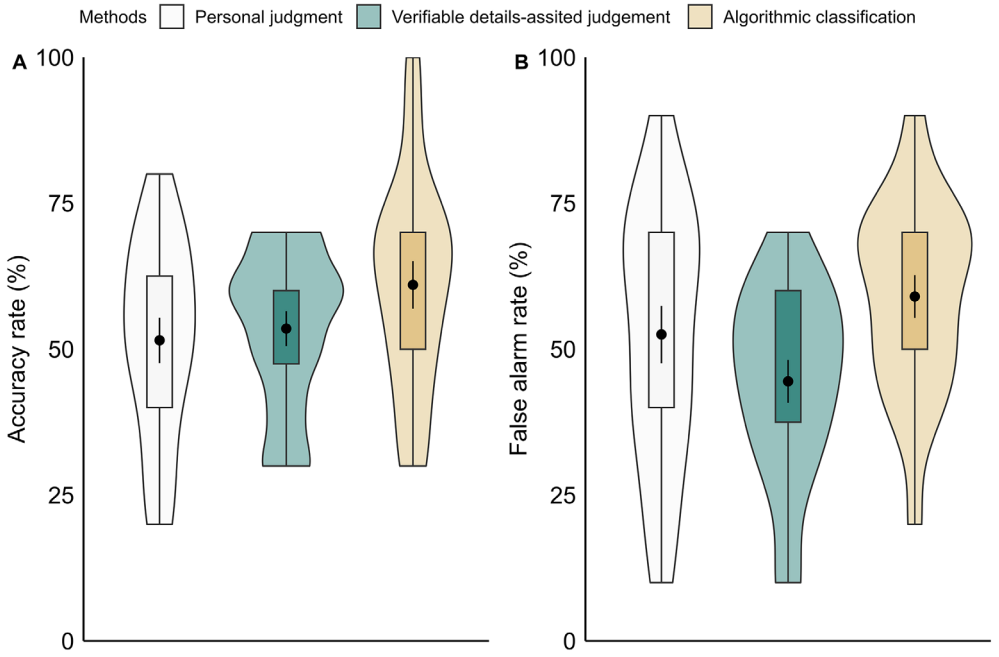

FIGURE 3 | Experiment 2. (A) Accuracy rates for methods; (B) False alarm rates for methods. Error bars represent standard errors of the mean. Note that Method D (consensus-based judgment protocol) is not shown in the figure as it was not included in the statistical comparison.

from random guessing ($p = 0.002$). The findings suggest that the verifiable detail classification algorithm performed better than random chance in detecting truthful and misleading statements.

The new anti-noise protocol based on judge majority consensus (method D: consensus-based judgment protocol) had an accuracy rate of 60% and a false alarm rate of 30%. This consensus-based judgment or anti-noise protocol mitigates the influence of individual noise and provides a more robust collective assessment of the statements. It should be noted that it is not possible to statistically compare the performance of the consensus-based judgment protocol (method D) with the other methods due to the data structure of this protocol.

The LME model for the accuracy rate (see Figure 3A) showed that the main effect of method was not statistically significant ($\chi^2_{(2)} = 3.765$, $p = 0.152$, $AIC_{RL} = 1.124$, $R^2_m = 0.062$, $R^2_c = 0.062$), suggesting that the inclusion of method did not substantially improve model fit compared to the null model. Thus, there was no significant difference between the accuracy rate of personal judgment (method A) ($M = 51.50$, $SD = 17.252$) compared to the verifiable details-assisted judgment (method B) (A–B comparison: $M = 53.50$, $SD = 13.485$, $t(42.1) = -0.384$, $p = 0.922$) and compared to the algorithmic classification (method C) (A–C comparison: $M = 61.\ 00$, $SD = 18.325$, $t = -1.822$, $p = 0.174$), as well as verifiable detail-assisted judgment (method B) ($M = 53.50$, $SD = 13.485$) compared to the algorithmic classification (method C) (B–C comparison: $M = 61.00$, $SD = 18.325$, $t(42.1) = -1.439$, $p = 0.330$).

In contrast, the LME model for the false alarm rate (see Figure 3B) showed that the main effect of method was statistically significant ($\chi^2_{(2)} = 6.206$, $p = 0.044$, $AIC_{RL} = 0.331$, $R^2_m = 0.099$, $R^2_c = 0.119$), indicating that the inclusion of method improved model fit relative to the null model. Planned comparisons showed no significant difference between the false alarm rate of the personal judgment (method A) ($M = 52.50$, $SD = 21.975$) compared to the verifiable detail-assisted judgment (method B) (A–B comparison: $M = 44.50$, $SD = 16.376$, $t(42.1) = 1.384$, $p = 0.358$) and compared to the algorithmic classification (method C) (A–C comparison: $M = 59.00$, $SD = 16.512$, $t(42.1) = -1.125$, $p = 0.504$). However, the false alarm rate of the verifiable detail-assisted judgment (method B) ($M = 44.50$, $SD = 16.376$) was significantly lower compared to the algorithmic classification (method C) (B–C comparison: $M = 59.00$, $SD = 16.512$, $t(42.1) = -2.509$, $p = 0.041$).

3.2.2 | Inter and Intra-Judge Variability Between Verifiable Details and Personal Judgment

The LME model for verifiable details showed that the main effect of the truthfulness of the statement was not statistically significant ($\chi^2_{(1)} = 2.830$, $p = 0.093$, $AIC_{RL} = 0.660$, $R^2_m = 0.006$, $R^2_c = 0.097$). Thus, there was no significant difference between the count of verifiable details between truthful ($M = 2.46$, $SD = 1.641$) and deceptive ($M = 2.20$, $SD = 1.665$, $t(381) = -1.683$, $p = 0.093$) statements.

Regarding individual judges who counted veracity details (method A: personal judgment), the rater fixed-effects model (ICC3) showed low and statistically significant consistency ($ICC3 = 0.36$, $p < 0.001$). This suggests that the same judge's judgments were inconsistent across truthful and deceptive statements. On the other hand, for all judges who performed truthfulness judgments, the random effects model for mean raters (ICC2k) showed high and statistically significant consistency ($ICC2k = 0.90$, $p < 0.001$). This indicates that judges were consistent in how they counted verifiable details for both truthful and deceptive statements.

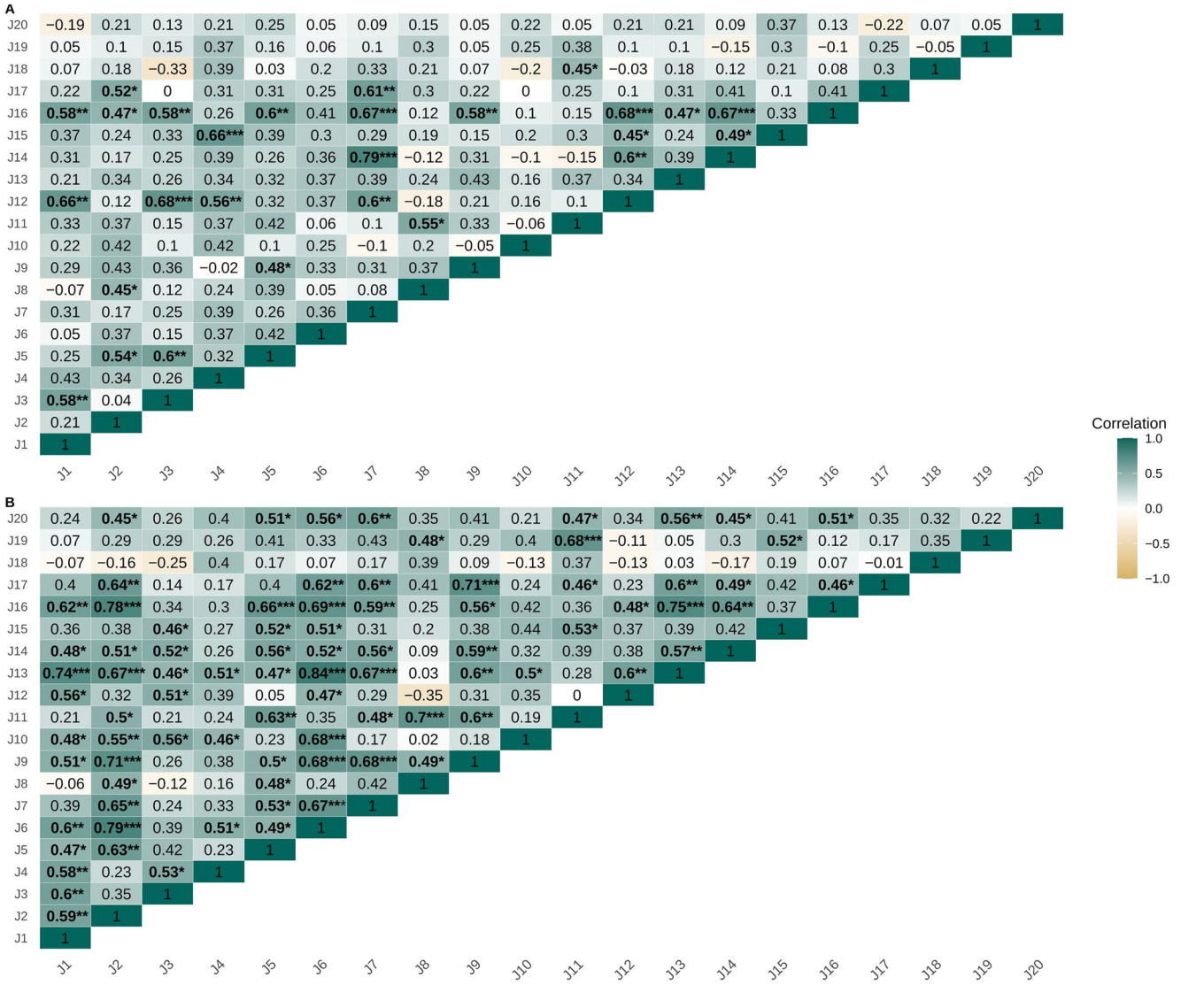

FIGURE 4 | Experiment 2. (A) Pearson's correlation matrix between judgments by the judges; (B) Pearson's correlation matrix between judges' verifiable detail counts. Bold values indicate statistical significance ($p < 0.05$*, $p < 0.01$**, $p < 0.001$***).

For individual judges who performed judgments after knowing the average verifiable details (method B: verifiable detail-assisted judgment) the rater fixed-effects model (ICC3) showed low and statistically significant consistency (ICC3 = 0.15, $p < 0.001$). This suggests that the same judge's judgments were inconsistent across truthful and deceptive statements. However, for all judges who performed these judgments, the random effects model for mean raters (ICC2k) showed high and statistically significant consistency (ICC2k = 0.78, $p < 0.001$). This indicates that judges with verifiable detail assistance also showed a high level of consistency in their judgments of both truthful and deceptive statements. As an illustration, in a deceptive statement, the number of verifiable details counted by judges ranged from 0 to 8 ($M = 2.65$, SD = 2.03). Interestingly, 55% ($n = 11$) of the judges considered the statement truthful, while 45% ($n = 9$) considered it deceptive. This highlights the substantial variability in personal judgments (method A).

For truthfulness judgments, the correlation matrix showed that the associations between judges were mostly low to moderate ($r_{mean} = 0.252$), with a few high correlations ($r = -0.33$ to 0.79). Only a few of these associations were statistically significant (see Figure 4A). Similarly, for counting verifiable details, associations between judges were mostly low to moderate ($r_{mean} = 0.383$), with a few high correlations ($r = -0.35$–0.84). However, a moderate number of these associations were statistically significant (see Figure 4B). These findings indicated a high degree of variability in their truthfulness judgments and verifiable detail counts.

3.3 | Discussion

Overall, the findings of experiment 1 were replicated in experiment 2. The results of experiment 1 showed that the VA performs better than chance (only in lie detection accuracy) when using the classification algorithm (method C: algorithmic classification) based on the average of verifiable details. However, individual human judgments (method A: personal judgment) and human judgments assisted by verifiable detail averaging (method B: verifiable detail-assisted judgment) did not show this same level of accuracy (they do not outperform chance).

Moreover, the count of verifiable details maintained a high variability when different statements were evaluated by the same judge, with no significant differences between truthful and deceptive statements. This reinforces the current limitations of the approach due to subjectivity in counting.

Beyond re-evaluating the noise introduced by detail counting, Experiment 2 also examined a second key source of noise: human judgment of truthfulness. This raised the question of which type of judgment might best reduce noise within the VA and improve deception detection. Although no significant differences were found between hit rates between human judgments, assisted and automated judgments, only the classification algorithm (method C) achieved an accuracy rate significantly above chance (50%) in detecting deception, whereas both unassisted (method A) and assisted (method B) human judgments did not. This suggests the classification algorithm (method C) may offer better practical performance than untrained human and assisted judgments. However, based on the average response of N independent judges, the anti-noise protocol (method D: consensus-based judgment protocol) matched the classification algorithm's overall accuracy (60%) and surpassed it in reducing false alarms (over 30% vs. over 50%). As in experiment 1, the classification algorithm (method C) effectively identified deceptive statements and misclassified many truthful statements. This replication strengthens the hypothesis that lacking clear criteria for counting verifiable details leads to either overestimation or underestimation of the average, thus impacting classification accuracy.

The results on variability and consistency of truthfulness and deception judgments showed high variability in individual judgments. This contributed to the overall low performance (i.e., low accuracy and high false alarm rates) and the hit rate not exceeding chance. On the other hand, the high group consistency suggests that the average group judgments across participants show relatively low variability across statements, indicating greater agreement and reliability at the group level. This reinforces the performance of the anti-noise protocol (method D) based on the average count, and the average truthfulness rating highlights that variability in individual judgments is a key source of noise in the VA.

Consequently, the classification algorithm based on verifiable details (method C) partially succeeds in reducing the two sources of noise of the VA: (a) subjectivity in counting verifiable details and (b) inconsistencies in judgments. False alarms remained random because they were frequent and did not show a consistent pattern of occurrence, suggesting that current methods may still allow for substantial error in classifying truthful statements. It is necessary to improve counting and human judgment to increase detection accuracy, mainly to decrease false alarms below random. In this regard, by averaging both noise sources, the anti-noise protocol (method D) outperformed the algorithm in reducing false alarms. These findings highlight the importance of minimizing subjectivity in all potential noise sources. Appropriate training (practice with feedback) for judges could be an effective strategy to achieve this goal. While the anti-noise protocol (method D) outperformed the algorithm in reducing false alarms, its performance was evaluated descriptively. Inferential statistical comparisons were not conducted because the consensus score yields a single binary classification per statement, based on majority voting across all raters. This structure produces a single accuracy value, with no within-condition variability, rendering standard inferential tests or bootstrapping techniques inapplicable. Future studies using resampling-based consensus methods or probabilistic scoring could address this limitation and allow for formal statistical comparisons between classification strategies. From an applied perspective, even small statistically significant increases in verifiable detail counts may translate into meaningful shifts in classification outcomes. In applied contexts that rely on cut-off-based decision rules (e.g., preliminary credibility assessments), such changes could affect whether a statement is flagged for further investigation. Although effect sizes were modest, even small increases in sensitivity to verifiability cues could yield meaningful operational benefits.

4 | General Discussion

Intuition-based lie detection is random, and it is easy to introduce biases toward truth or lies (Bogaard et al. 2016b). For example, manipulating judges' expectations through instructions about the proportion of lies can create a bias toward truthfulness or increase false alarms. This occurs when the default expectation in many studies (i.e., 50% of statements being lies) is altered, for instance, informing judges that only 20% of statements are lies may bias them toward classifying more statements as truthful, while telling them that 80% are lies may lead to an increase in false alarms. Participants are unaware of their lie-detection ability or credulity until their performance is explicitly measured or compared against objective outcomes. A simple method to overcome intuition-based lie detection is the VA (Nahari et al. 2014a, 2014b). Despite its simplicity, the method still carries the risk of variability in identifying verifiable details and judging truthfulness, both of which introduce potential noise sources. Experiments 1 and 2 explored these two noise sources and consistently replicated the effectiveness of the classification algorithm based on verifiable details to identify deceptive statements. In this discussion, we refer to four approaches tested in our experiments, particularly in Experiment 2: (i) personal judgments (based solely on each judge's own count of verifiable details; method A in Experiment 2), (ii) verifiable detail-assisted judgments (where judges were provided with the average verifiable detail count per statement; method B), (iii) algorithmic classification (which uses a fixed cut-off based on the overall average of verifiable details; method C), and (iv) a consensus-based (anti-noise) protocol that aggregates independent judgments to enhance reliability and reduce noise (method D).

The results obtained in both experiments highlight the potential and limitations of VA for deception detection. The algorithmic classification (method C) based on verifiable details showed significantly higher accuracy than chance. In contrast, personal judgments (method A) and verifiable detail-assisted judgments (method B) did not meet this threshold. However, the high false alarm rate and variability associated with human judgment highlight critical challenges in the practical application of this approach. Both experiments also showed that aggregation of group judgments using an anti-noise protocol can reduce the impact of these limitations.

Specifically, in Experiment 1, it was observed that judges tend to count more unverifiable details in truthful statements than in deceptive ones. At the same time, no significant differences were found in the count of verifiable details between the two statement types. This suggests that focusing exclusively on verifiable details, rather than using the ratio of verifiable to unverifiable, provides a more reliable indicator. The high variability in individual counts and inconsistency between judges were identified as primary noise sources contributing to the elevated false alarm rate. Experiment 2 confirmed these findings and provided further evidence for two key noise sources: subjectivity in identifying verifiable details and inconsistency in veracity judgments. This approach directly responds to previously identified limitations in applying the VA, particularly the variability in untrained or non-standardized assessments reported by Palena et al. (2021). Rather than treating this issue as novel, we sought to develop and test methods to reduce judgment inconsistency and mitigate the illusion of agreement arising from apparent agreement among raters. Although both the algorithmic classification (method C) and the consensus-based protocol (method D) reached similar accuracy rates, only the latter substantially reduced false alarms (30%), thus offering greater practical applicability.

One of the study's main strengths was replicating the findings between experiments. This reinforces the robustness of the conclusions. In addition, the combined use of human-based judgments (methods A and B) and algorithmic and consensus-based classification (methods C and D) allowed the evaluation of various strategies to address the two sources of noise in the VA. The findings of this study highlight the need to address the inherent subjectivity of VA. Although the classification algorithm used in this study relies on a simple cut-off rule based on the average number of verifiable details, this approach was deliberately selected to prioritize interpretability and practical feasibility in non-computational or time-sensitive environments. While not optimized, this rule provided better-than-chance accuracy. It significantly outperformed personal judgment (method A), establishing a useful baseline. This can serve as a reference against which more sophisticated models, such as logistic regression or decision trees, can be evaluated in future research. In this vein, replicating these findings in a sample of "super analyst" participants based on a specific psychological profile (Kahneman et al. 2021) could also offer valuable insights. Finally, investigating the applicability of these findings in practical contexts, such as police interviews or legal evaluations, could significantly expand the impact of VA.

To summarize, this study highlights both the strengths and limitations of the VA. While the algorithmic classification (method C) showed clear advantages in accuracy, false alarms remain a critical barrier to its implementation, and the average-based anti-noise protocol offers a promising alternative to improve stability and reduce judgment inconsistencies. Furthermore, future research could explore the use of advanced algorithms, such as machine learning models, to further improve accuracy and reduce false alarms. Another important limitation is the absence of formative training for the judges, which could help explain the high variability observed in verifiable detail counts. However, this design choice was intentional and theoretically grounded. According to Kahneman et al. (2021), the first step in addressing judgmental noise is to evaluate the extent to which variability is due to incompetence or lack of task understanding. Establishing this baseline is essential before designing interventions such as training or feedback. Furthermore, in many real-world applications, such as forensic or operational settings, decision-makers often do not receive feedback on the accuracy of their judgments. Therefore, developing noise-reduction strategies that remain effective under feedback-free conditions is critical for improving the practical utility of the VA. Our findings provide an empirical starting point for these efforts. In this sense, this study emphasizes that, beyond the content of the statements themselves, variability in human evaluators represents a critical and previously underexplored source of noise in the application of the VA.

Another methodological limitation concerns using linear mixed-effects models that only included random intercepts for participants, without incorporating random effects for statements or random slopes. This decision was made to preserve model parsimony, given our modest sample size ($n = 18$ for Experiment 1 and $n = 20$ for Experiment 2), which limits the stability of more complex models with multiple parameters (Barr et al. 2013; Matuschek et al. 2017). However, including random intercepts for statements, or even random slopes for participants, could provide additional insight into variability in cue sensitivity across individuals and statements. As Hudson et al. (2020) noted, modeling this variability may reveal how some individuals are more attuned to certain cues of veracity than others. Future studies with larger samples could use these more advanced modeling approaches to explore individual and item-level dynamics in greater depth.

A further consideration is the reliance on average-based algorithms, which introduces challenges related to interpretability and practical implementation. Our study's relatively modest number of judges suggests that judgmental noise may not be the sole contributor to variability. Ambiguity at the stimulus level or uncertainty in the coding process may also contribute, such as unclear statements or inconsistencies in how raters interpret and code verifiability. Suppose the VA requires algorithmic or group corrections to achieve adequate performance. In that case, this raises important concerns about whether the VA, in its current form, can function effectively as a practical, stand-alone lie detection tool. This approach appears to depend on structured training, judge independence, practice with feedback, and a well-defined minimum number of independent raters (N), which may vary depending on the context. Palena et al. (2021) noted that the VA performs better in high-stakes environments such as airports or crime scenes than in occupational or malingering contexts.

Our results suggest that consensus-based ratings may outperform individual ones, while classification using cut-off algorithms also shows promise. It remains an open empirical question whether these strategies are interchangeable, complementary, or more appropriate at different stages or in different applications. For example, one approach may be favored depending on whether prior decision data or feedback-informed classification histories are available. Although consensus-based judgments (method D) and algorithmic cut-off classifications (method C) effectively reduce noise, they rely on fundamentally

different principles. We therefore recognize that these approaches are not interchangeable. Consensus-based methods (method D: consensus-based judgment protocol) reflect the aggregation of human judgments and draw on the wisdom of crowds. In contrast, algorithmic rules (method C: algorithmic classification) apply fixed statistical thresholds. Each offers distinct trade-offs: consensus may provide greater nuance but requires more human input, whereas algorithms are more scalable but potentially less adaptive. In applied contexts, the choice between them will likely depend on available resources, time constraints, and the feasibility of deploying trained evaluators or automated tools. Building on this distinction, future research should replicate Experiments 1 and 2 with trained judges, including feedback-based practice and selection based on relevant cognitive profiles. It is also essential to systematically examine the minimum number of independent raters required for a reliable classification.

Taken together, our findings suggest that improving the VA requires refining how statement content is assessed and systematically addressing the cognitive variability of those who apply it. Although both the algorithmic classification (method C) and the consensus-based judgment protocol (anti-noise protocol; method D) proved helpful in mitigating noise sources, further efforts are needed to improve accuracy and reduce false alarms. These findings underscore the importance of developing robust and applicable methods to address the inherent complexity of deception detection in real-world settings. In sum, while the VA is promising, it will only reach its full potential through a dual refinement: enhancing procedural mechanisms and ensuring cognitive consistency among its evaluators.

Author Contributions

Germán Gálvez-García: conceptualization, investigation, funding acquisition, writing – original draft, methodology, validation, visualization, writing – review and editing, software, formal analysis, project administration, data curation, supervision, resources. **Patricio Mena-Chamorro:** conceptualization, investigation, methodology, validation, visualization, writing – original draft, writing – review and editing, software, formal analysis, data curation, supervision, funding acquisition. **Alejandro Moliné:** conceptualization, investigation, writing – original draft, methodology, software, data curation. **Tomás Espinoza-Palavicino:** data curation, software, formal analysis, project administration, writing – review and editing, visualization, validation, methodology, conceptualization, investigation, funding acquisition, writing – original draft. **Oscar Iborra:** conceptualization, investigation, writing – original draft, methodology, software, data curation. **Emilio Gómez-Milán:** conceptualization, investigation, funding acquisition, writing – original draft, methodology, validation, visualization, writing – review and editing, software, formal analysis, project administration, data curation, supervision, resources.

Acknowledgments

This research was partially supported by FONDECYT project 1230431 from ANID granted to Germán Gálvez-García and two PhD scholarships from ANID (21210298 and 21231269) for Patricio Mena-Chamorro and Tomás Espinoza-Palavicino respectively).

Ethics Statement

This study was approved by the Research Ethics Committee of the University of Granada (registration number 938/CEIH/2019).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are openly available in OSF at https://osf.io/j45ns/?view_only=51c7ab1e7fe04407b5da1377df8013e7.

References

Barr, D. J., R. Levy, C. Scheepers, and H. J. Tily. 2013. "Random Effects Structure for Confirmatory Hypothesis Testing: Keep It Maximal." *Journal of Memory and Language* 68, no. 3: 255–278. https://doi.org/10.1016/j.jml.2012.11.001.

Bates, D., M. Mächler, B. Bolker, and S. Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67, no. 1: 1–48. https://doi.org/10.18637/jss.v067.i01.

Bogaard, G., K. Colwell, and S. Crans. 2019. "Using the Reality Interview Improves the Accuracy of the Criteria-Based Content Analysis and Reality Monitoring." *Applied Cognitive Psychology* 33, no. 6: 1018–1031. https://doi.org/10.1002/acp.3537.

Bogaard, G., E. H. Meijer, A. Vrij, and H. Merckelbach. 2016a. "Scientific Content Analysis (SCAN) Cannot Distinguish Between Truthful and Fabricated Accounts of a Negative Event." *Frontiers in Psychology* 7: 243. https://doi.org/10.3389/fpsyg.2016.00243.

Bogaard, G., E. H. Meijer, A. Vrij, and H. Merckelbach. 2016b. "Strong, but Wrong: Lay People's and Police Officers' Beliefs About Verbal and Nonverbal Cues to Deception." *PLoS One* 11, no. 6: e0156615. https://doi.org/10.1371/journal.pone.0156615.

Bogaard, G., E. H. Meijer, and I. Van der Plas. 2020. "A Model Statement Does Not Enhance the Verifiability Approach." *Applied Cognitive Psychology* 34, no. 1: 96–105. https://doi.org/10.1002/acp.3596.

Bond, C. F., and B. M. DePaulo. 2008. "Individual Differences in Judging Deception: Accuracy and Bias." *Psychological Bulletin* 134, no. 4: 477–492. https://doi.org/10.1037/0033-2909.134.4.477.

Boskovic, I., G. Bogaard, H. Merckelbach, A. Vrij, and L. Hope. 2017. "The Verifiability Approach to Detection of Malingered Physical Symptoms." *Psychology, Crime & Law* 23, no. 8: 717–729. https://doi.org/10.1080/1068316X.2017.1302585.

Cole, T. 2001. "Lying to the One You Love: The Use of Deception in Romantic Relationships." *Journal of Social and Personal Relationships* 18: 107–129. https://doi.org/10.1177/0265407501181005.

Deeb, H., A. Vrij, S. Leal, and S. Mann. 2021. "Combining the Model Statement and the Sketching While Narrating Interview Techniques to Elicit Information and Detect Lies in Multiple Interviews." *Applied Cognitive Psychology* 35, no. 6: 1478–1491. https://doi.org/10.1002/acp.3880.

Denault, V. 2020. "Misconceptions About Nonverbal Cues to Deception: A Covert Threat to the Justice System?" *Frontiers in Psychology* 11: 573460. https://doi.org/10.3389/fpsyg.2020.573460.

Denault, V., P. Plusquellec, L. M. Jupe, et al. 2020. "The Analysis of Nonverbal Communication: The Dangers of Pseudoscience in Security and Justice Contexts." *Anuario de Psicología Jurídica* 30, no. 1: 1–12. https://doi.org/10.5093/apj2019a9.

DePaulo, B. M., D. A. Kashy, S. E. Kirkendol, M. M. Wyer, and J. A. Epstein. 1996. "Lying in Everyday Life." *Journal of Personality and Social Psychology* 70, no. 5: 979–995. https://doi.org/10.1037/0022-3514.70.5.979.

Espinoza-Palavicino, T., P. Mena-Chamorro, J. Albayay, A. Doussoulin, and G. Gálvez-García. 2023. "The Use of Transcutaneous Vagal Nerve Stimulation as an Effective Countermeasure for Simulator Adaptation

Syndrome." *Applied Ergonomics* 107: 103921. https://doi.org/10.1016/j.apergo.2022.103921.

Faul, F., E. Erdfelder, A. Buchner, and A. Lang. 2009. "Statistical Power Analyses Using G* Power 3.1: Tests for Correlation and Regression Analyses." *Behavior Research Methods* 41, no. 4: 1149–1160.

Gálvez-García, G., J. Fernández-Gómez, C. Bascour-Sandoval, et al. 2021. "A Trifactorial Model of Detection of Deception Using Thermography." *Psychology, Crime & Law* 27, no. 5: 405–426. https://doi.org/10.1037/0022-3514.70.5.979.

Gálvez-García, G., P. Mena-Chamorro, T. Espinoza-Palavicino, T. Romero-Arias, M. Barramuño-Medina, and C. Bascour-Sandoval. 2024. "Mixing Transcutaneous Vagal Nerve Stimulation and Galvanic Cutaneous Stimulation to Decrease Simulator Adaptation Syndrome." *Frontiers in Psychology* 15: 1476021. https://doi.org/10.3389/fpsyg.2024.1476021.

Granhag, P. A., and M. Hartwig. 2015. "The Strategic Use of Evidence (SUE) Technique: A Conceptual Overview." In *Detecting Deception: Current Challenges and Cognitive Approaches*, edited by P. A. Granhag, A. Vrij, and B. Verschuere, 231–251. Wiley Blackwell.

Hartwig, M., and C. F. Bond Jr. 2011. "Why Do Lie-Catchers Fail? A Lens Model Meta-Analysis of Human Lie Judgments." *Psychological Bulletin* 137, no. 4: 643. https://doi.org/10.1037/a0023589.

Hartwig, M., and C. F. Bond Jr. 2014. "Lie Detection From Multiple Cues: A Meta-Analysis." *Applied Cognitive Psychology* 28: 661–676. https://doi.org/10.1002/acp.3052.

Harvey, A. C., A. Vrij, S. Leal, M. Lafferty, and G. Nahari. 2017. "Insurance Based Lie Detection: Enhancing the Verifiability Approach With a Model Statement Component." *Acta Psychologica* 174: 1–8. https://doi.org/10.1016/j.actpsy.2017.01.001.

Hauch, V., S. L. Sporer, S. W. Michael, and C. A. Meissner. 2016. "Does Training Improve the Detection of Deception? A Meta-Analysis." *Communication Research* 43: 283–343. https://doi.org/10.1177/0093650214534974.

Hudson, C. A., A. Vrij, L. Akehurst, L. Hope, and L. P. Satchell. 2020. "Veracity is in the Eye of the Beholder: A Lens Model Examination of Consistency and Deception." *Applied Cognitive Psychology* 34, no. 5: 996–1004.

Jensen, L. A., J. J. Arnett, S. S. Feldman, and E. Cauffman. 2004. "The Right to Do Wrong: Lying to Parents Among Adolescents and Emerging Adults." *Journal of Youth and Adolescence* 33: 101–112. https://doi.org/10.1023/B:JOYO.0000013422.48100.5a.

Jupe, L. M., S. Leal, A. Vrij, and G. Nahari. 2017. "Applying the Verifiability Approach in an International Airport Setting." *Psychology, Crime & Law* 23, no. 8: 812–825. https://doi.org/10.1080/1068316X.2017.1327584.

Kahneman, D., O. Sibony, and C. R. Sunstein. 2021. *Noise: A Flaw in Human Judgment*. Hachette UK.

Köhnken, G., A. L. Manzanero, and M. T. Scott. 2015. "Análisis de la validez de las declaraciones: mitos y limitaciones." *Anuario de Psicología Jurídica* 25: 13–19. https://doi.org/10.1016/j.apj.2015.01.004.

Lefcheck, J. S. 2016. "piecewiseSEM: Piecewise Structural Equation Modeling in R for Ecology, Evolution, and Systematics." *Methods in Ecology and Evolution* 7, no. 5: 573–579. https://doi.org/10.1111/2041-210X.12512.

Lenth, R. V. 2023. "emmeans: Estimated Marginal Means, Aka Least-Squares Means (versión 1.8.3)." https://cran.r-project.org/package=emmeans.

Masip, J. 2017. "Deception Detection: State of the Art and Future Prospects." *Psicothema* 2: 149–159. https://doi.org/10.7334/psicothema2017.34.

Matuschek, H., R. Kliegl, S. Vasishth, H. Baayen, and D. Bates. 2017. "Balancing Type I Error and Power in Linear Mixed Models." *Journal of Memory and Language* 94: 305–315. https://doi.org/10.1016/j.jml.2017.01.001.

McElreath, R. 2016. *Statistical Rethinking: A Bayesian Course With Examples in R and Stan*. CRC Press.

Nahari, G. 2018. "The Applicability of the Verifiability Approach to the Real World." In *Detecting Concealed Information and Deception*, 329–349. Academic Press.

Nahari, G., A. Vrij, and R. P. Fisher. 2014a. "Exploiting Liars' Verbal Strategies by Examining the Verifiability of Details." *Legal and Criminological Psychology* 19, no. 2: 227–239. https://doi.org/10.1111/j.2044-8333.2012.02069.x.

Nahari, G., A. Vrij, and R. P. Fisher. 2014b. "The Verifiability Approach: Countermeasures Facilitate Its Ability to Discriminate Between Truths and Lies." *Applied Cognitive Psychology* 28, no. 1: 122–128. https://doi.org/10.1002/acp.2974.

Palena, N., L. Caso, A. Vrij, and G. Nahari. 2021. "The Verifiability Approach: A Meta-Analysis." *Journal of Applied Research in Memory and Cognition* 10, no. 1: 155–166. https://doi.org/10.1037/h0101785.

RStudio Team. 2020. *RStudio: Integrated Development for R*. RStudio, PBC.

Serota, K. B., T. R. Levine, and F. J. Boster. 2010. "The Prevalence of Lying in America: Three Studies of Self-Reported Lies." *Human Communication Research* 36: 2–25. https://doi.org/10.1111/j.1468-2958.2009.01366.x.

Sporer, S. L., and B. Schwandt. 2007. "Moderators of Nonverbal Indicators of Deception: A Meta-Analytic Synthesis." *Psychology, Public Policy, and Law* 13, no. 1: 1–34.

Strömwall, L. A., M. Hartwig, and P. A. Granhag. 2006. "To Act Truthfully: Nonverbal Behaviour and Strategies During a Police Interrogation." *Psychology, Crime & Law* 12, no. 2: 207–219. https://doi.org/10.1080/10683160512331331328.

Vernham, Z., A. Vrij, G. Nahari, et al. 2020. "Applying the Verifiability Approach to Deception Detection in Alibi Witness Situations." *Acta Psychologica* 204: 103020. https://doi.org/10.1016/j.actpsy.2020.103020.

Verschuere, B., G. Bogaard, and E. Meijer. 2021. "Discriminating Deceptive From Truthful Statements Using the Verifiability Approach: A Meta-Analysis." *Applied Cognitive Psychology* 35, no. 2: 374–384. https://doi.org/10.1002/acp.3775.

Verschuere, B., C.-C. Lin, S. Huismann, et al. 2023. "The Use-The-Best Heuristic Facilitates Deception Detection." *Nature Human Behaviour* 7, no. 5: 718–728. https://doi.org/10.1038/s41562-023-01556-2.

Verschuere, B., M. Schutte, S. van Opzeeland, and I. Kool. 2021. "The Verifiability Approach to Deception Detection: A Preregistered Direct Replication of the Information Protocol Condition of Nahari, Vrij, and Fisher (2014b)." *Applied Cognitive Psychology* 35, no. 1: 308–316. https://doi.org/10.1002/acp.3769.

Vrij, A. 2019. "Deception and Truth Detection When Analyzing Nonverbal and Verbal Cues." *Applied Cognitive Psychology* 33, no. 2: 160–167. https://doi.org/10.1002/acp.3457.

Vrij, A., R. P. Fisher, H. Blank, S. Leal, and S. Mann. 2016. "A Cognitive Approach to Elicit Verbal and Nonverbal Cues to Deceit." In *Cheating, Corruption, and Concealment: The Roots of Dishonesty*, edited by J.-W. van Prooijen and P. A. M. van Lange, 284–302. Cambridge University Press.

Vrij, A., P. A. Granhag, T. Ashkenazi, G. Ganis, S. Leal, and R. P. Fisher. 2022. "Verbal Lie Detection: Its Past, Present and Future." *Brain Sciences* 12, no. 12: 1644. https://doi.org/10.3390/brainsci12121644.

Vrij, A., S. Leal, and R. P. Fisher. 2018. "Verbal Deception and the Model Statement as a Lie Detection Tool." *Frontiers in Psychiatry* 9: 492. https://doi.org/10.3389/fpsyt.2018.00492.

Vrij, A., S. Mann, S. Leal, and R. Fisher. 2010. "'Look Into My Eyes': Can an Instruction to Maintain Eye Contact Facilitate Lie Detection?" *Psychology, Crime & Law* 16, no. 4: 327–348. https://doi.org/10.1080/10683160902740633.

Vrij, A., and G. Nahari. 2019. "The Verifiability Approach." In *Evidence-Based Investigative Interviewing: Applying Cognitive Principles*, edited by J. J. Dickinson, N. S. Compo, R. N. Carol, B. L. Schwartz, and M. R. McCauley, 116–133. Routledge/Taylor & Francis Group.