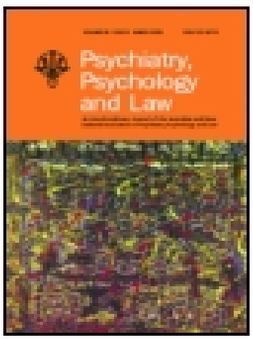

Psychiatry, Psychology and Law

ISSN: 1321-8719 (Print) 1934-1687 (Online) Journal homepage: https://www.tandfonline.com/loi/tppl20

A within-statement baseline comparison for detecting lies

Brianna L. Verigin, Ewout H. Meijer & Aldert Vrij

To cite this article: Brianna L. Verigin, Ewout H. Meijer & Aldert Vrij (2020): A within-statement baseline comparison for detecting lies, Psychiatry, Psychology and Law

To link to this article: https://doi.org/10.1080/13218719.2020.1767712

Published online: 02 Jun 2020.

Psychiatry, Psychology and Law, 2020
Vol. 0, No. 0, 1–10, https://dx.doi.org/10.1080/13218719.2020.1767712

A within-statement baseline comparison for detecting lies

Brianna L. Verigin[a,b], Ewout H. Meijer[a] and Aldert Vrij[b]

[a]*Forensic Psychology Section, Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, Netherlands;* [b]*Department of Psychology, University of Portsmouth, Portsmouth, UK*

To make veracity judgements in individual cases, practitioners may rely on baselining. That is, they may evaluate a statement relative to a baseline statement that is known to be truthful. We investigated whether a within-statement verbal baseline comparison could enhance discriminatory accuracy. Participants ($n = 148$) read an alibi statement of a mock suspect and provided a veracity judgement regarding a critical two-hour period within the alibi statement. This critical element was either deceptive or truthful and was embedded into an otherwise truthful story. Half of the participants received additional instructions to use the surrounding truthful elements of the statement as a baseline. Instructing participants to make a within-statement baseline comparison did not improve the accuracy of credibility assessments.

Key words: baseline technique; comparable truth; deception detection; interviewing techniques; veracity assessment; within-statement comparison; within-subjects lie detection.

Introduction

Deception researchers typically report their results as average scores from groups of participants. This research shows that, on average, liars' statements are less richly detailed than those of their truth-telling counterparts (e.g. Amado, Arce, Fariña, & Vilarino, 2016; DePaulo et al., 2003; Luke, 2019). Legal practitioners, in contrast, are rarely interested in such group averages. They need to know whether an interviewee in the case at hand is being deceptive or honest. However, group-derived estimates do not always reliably generalize to individual cases (Fisher, Medaglia, & Jeronimus, 2018; Faigman, Monahan, & Slobogin, 2014).

One option that facilitates decisions at the individual level is to include a within-individual comparison (see Vrij, 2016, for a discussion). One such method, reportedly used in practice by some police (Ewens, Vrij, Jang, & Jo, 2014; Frank, Yarbrough, & Ekman, 2006; Inbau, Reid, Buckley, & Jayne, 2013; U.S. Department of the Army, 2006) is the baseline technique. With this technique, interviewers evaluate an interviewee's 'statement of interest' (i.e. part of the statement for which veracity is being assessed) relative to a baseline statement (i.e. part of the interview that is known to be

Correspondence: Brianna L. Verigin, Faculty of Psychology and Neuroscience, Maastricht University, Universiteitssingel 40, 6229 ER Maastricht, Netherlands. Email: brianna.verigin@maastrichtuniversity.nl

truthful). Deception is then determined by looking for deviations from this established baseline. Moreover, 71% of experienced human intelligence interviewers reported to rely on deviations from baseline to detect deception (Russano, Narchet, Kleinman, & Meissner, 2014).

A reason to believe in the efficacy of baselining as a lie detection technique derives from early research on the relationship between familiarity and deception. A handful of studies from the late 20th century examined how the level of familiarity between a liar and an observer affects lie detection outcomes (e.g. Brandt, Miller, & Hocking, 1980a, 1980b, 1982; Comadena, 1982; Ekman & Friesen, 1974; Feeley, deTurck, & Young, 1995; Hayano, 1980; McCornack & Parks, 1986). Collectively, this research showed that veracity judgements were most accurate when observers had the opportunity to become familiar with the respondents' truthful communication style. For example, Feeley et al. (1995) had participants judge the veracity of truthful and deceptive communicators after viewing between zero and four exposures of the sender. They found a positive linear relationship between the amount of familiarity with the sender and judges' accuracy. In their meta-analysis, Bond and DePaulo (2006) compared deception detection accuracy between judges who had, versus those who had not, been previously exposed to the individual they were evaluating. Their results showed that when judges had been previously exposed to a target, this baseline exposure or baseline familiarity resulted in a small yet significant increase in detection accuracy from 52% to 56%. These results support the role of baseline familiarity – that is, when the sender is familiar to the receiver, such as in personal relationships.

In police interviews, interviewer and target are more likely to be strangers, and another option is to use part of the same interview as the baseline statement. Ewens and colleagues (2014) examined the behavioural patterns of interviewees in response to an initial non-threatening 'small-talk' baseline question ('You just read and signed an informed consent form, could you please tell me what you remember about it and what it said') compared to their behaviour in response to investigative questions, for which they knew their veracity would be assessed. The results indicated no effect of this baseline: both truthful and deceptive interviewees behaved equally different between the small-talk baseline and investigative part of the interview. Palena, Caso, Vrij, and Orthey (2018) examined two types of baseline: an initial small-talk baseline and a comparable truth baseline (i.e. a set of questions designed to be comparable with investigative questions in terms of, for example, content, stakes and time-frame; Ewens et al., 2014; Vrij, 2008). They compared similarities in participants' nonverbal and verbal behaviours when responding to baseline and investigative questions, using the two types of baseline. They found that liars and truth-tellers in the small-talk baseline condition did not differ in their level of similarity between the baseline and investigative questions, adding further evidence to the ineffectiveness of this approach. Their results did, however, reveal that truth-tellers showed significantly more similarity than liars in the comparable baseline condition, though only in terms of spatial details.

Two previous studies examined the effect of verbal baselining on observers' accuracy rates. Caso, Palena, Carlessi, and Vrij (2019) tested police officers' ability to assess credibility when provided with a comparable truth baseline compared to when no baseline was provided. No differences were found for total and truth accuracy between conditions, although observers who made a baseline comparison did obtain higher lie detection accuracy rates. Caso, Palena, Vrij, and Gnisci (2019) found more promising results. These authors looked at the effects of small-talk and a comparable truth baseline on laypeople's deception detection accuracy. This study

revealed that (a) participants in the comparable truth condition outperformed those in the small-talk condition in terms of total accuracy rates, and (b) only observers who used a comparable truth baseline performed significantly better than chance levels in their total accuracy for distinguishing truth-tellers from liars.

Taken together, previous research on baselining shows that to enhance diagnostic accuracy, a comparable truth baseline should be used. That is, the baseline statement must be equivalent to the statement of interest in terms of content, time-frame, stakes, cognitive and emotional involvement, and questioning context (Caso, Palena, Vrij, et al., 2019; Ewens et al., 2014; Palena, Caso, Vrij, & Orthey, 2018; Vrij, 2008). Despite the importance of the baseline statement being equivalent to the target portion of the statement, each of the previous studies compared the effect of an initial, separate baseline statement to a target portion of an investigative interview. Investigating whether a baseline statement could be derived from parts of the interviewee's statement could have important implications for practitioners who may be inclined to draw such comparisons between corroborated and uncorroborated portions of an interviewee's account.

The objective of the present study was to investigate whether introducing a within-statement baseline comparison could improve the accuracy of participants' veracity judgements. Participants read the alibi statement of a mock suspect and provided a veracity judgement regarding a critical two-hour period within the alibi statement. This critical element was either deceptive or truthful and was embedded into an otherwise truthful story. We examined whether providing an instruction to utilize a comparable baseline (i.e. informing participants that all information, with the exception of the critical element, has been confirmed to be truthful) could enhance participants' detection accuracy. We hypothesized that participants who received the baseline instruction would have more accurate veracity judgements than participants who did not receive the baseline instruction.

Method

Participants

The sample consisted of 148 adult participants (120 females; 28 males) between the ages of 17 and 45 years (M_{age} = 20.53 years, SD = 3.17). Our sample size was calculated prior to data collection by multiplying the total number of statements ($n = 74$) by two, ensuring that each statement was evaluated twice. Given this sample size, and an α of .05, we had an 85.6% chance of rejecting the null hypothesis if there was a medium effect size (f = 0.25; Cohen, 1988). Only participants who were proficient in reading and writing English were eligible for the study. They were compensated with either course credit or a €5 voucher. The study was approved by the standing ethical committee. The study was pre-registered and approved via the Open Science Framework (http://j.mp/2IjvL51).

Statements

The statements that participants evaluated were previously collected by the authors (Verigin, Meijer, & Vrij, 2020). These statements represent accounts provided by student research participants who were instructed to provide oral alibi statements to convince an interviewer that they were innocent of a hypothetical crime. For the present study, we incorporated the statements that were entirely truthful recollections of an interviewee's events on a particular day ($n = 37$), and the statements that were truthful accounts containing a lie from 1.00pm to 3.00pm ($n = 37$).[1]

[1] For the latter group, interviewees truthfully reported their events on the day in question, before 1.00pm and after 3.00pm; however, during the critical element (i.e. between 1.00pm and 3.00pm) they were instructed to fabricate a particular activity. Thus, participants in the current study assessed

one transcript that contained a critical element that was either deceptive (i.e. embedded into an otherwise truthful account) or truthful (i.e. part of an entirely truthful account). Each of the 74 statements was evaluated twice by two independent participants.

Ground truth

We attempted to establish partial ground truth of the statements by asking participants to self-report the truthfulness of both elements of their statement (on a scale of 1 to 10, 1 being not at all truthful and 10 being completely truthful). Truth-tellers reported that both their general alibi ($M = 9.32$, $SD = 0.88$) and the critical element ($M = 9.59$, $SD = 0.90$) were almost completely truthful. Those who provided the embedded lie reported that their general alibi was almost entirely truthful ($M = 7.92$, $SD = 2.45$) whereas the critical element was mostly deceptive ($M = 2.62$, $SD = 2.48$). Thus, interviewees appeared to have largely conformed to the instructions they received across conditions.

Design

The experiment followed a between-subjects factorial design: 2 (baseline instruction: present vs. absent) × 2 (veracity of the critical element: truth vs. lie). Participants were randomly assigned to one of the four conditions. To mimic real-life investigations in which investigators typically only have one statement to assess, each participant judged only one statement. The dependent measure was the accuracy of participants' veracity judgements. Two accuracy scores were created by recoding participants' binary and Likert scale truth–lie judgements with the ground truth of the veracity of the critical element.

Procedure

Participants arrived at the lab and provided informed consent. Afterwards, they received a detailed instruction letter (see Appendix) explaining that their task was to imagine themselves in the role of a police detective who was investigating a violent burglary that had occurred recently. Participants were told that the suspect was interviewed by police and had provided an alibi statement for the entire day in question, from morning to evening. They were informed that the critical element of the alibi was from 1.00pm to 3.00pm on this day. The critical element within each transcript was highlighted yellow to ensure this was clearly understood. Participants were instructed to read the entire statement carefully, but to make an assessment regarding the veracity of only the highlighted critical element. All participants were told that it was important to make the correct decision because it would earn them a chance to win €50 from a raffle draw.

Participants who were assigned to the baseline-present condition received additional instructions prior to reading the transcript. They were informed that as the lead investigator, they had access to other sources of information for the case, and this collateral evidence confirmed that the 'general' alibi statement, before 1.00pm and after 3.00pm, was truthful (participants were not actually provided this collateral evidence). Participants were instructed to use this knowledge to compare the 'general' portion of the interviewees' alibi to the 'critical element from 1.00pm to 3.00pm'. They were asked to try to identify any patterns or changes in the verbal content between the general alibi and the critical element that may indicate how credible the suspect's account was during the highlighted critical element.

After reading the instructions, all participants received one written transcript of a suspect's alibi statement, and they were given up to 10 minutes to read it. We used written transcripts to allow for highlighting the critical element in yellow. After reading the transcript, participants were prompted to first provide a binary veracity judgement (lie or truth) regarding the highlighted critical element. Then, they rated their veracity judgement on a 7-point

Table 1. Descriptive statistics of binary and Likert veracity judgements across conditions.

	Baseline instruction		No baseline instruction	
	Deceptive element	Truthful element	Deceptive element	Truthful element
Binary judgement	.73 (.45)	.38 (.49)	.59 (.50)	.43 (.50)
Likert judgement	4.62 (1.44)	3.92 (1.57)	3.89 (1.49)	4.16 (1.57)

Note: Means; standard deviations in parentheses. The binary judgements have been recoded for accuracy, whereas the Likert scores are in their original form.

Likert scale (1 = *completely truthful* to 7 = *completely deceptive*). Subsequently, participants were prompted (a) to provide an open-ended description of the verbal cues they used to form their veracity judgement and (b) to select veracity cues from a predetermined list. To preserve manuscript length, the coding and analyses of participants' veracity cues are reported in the Supplementary Materials. Once completed, participants responded to a short questionnaire that included a motivation check, general study experience questions and demographics information (i.e. age, sex, race, native language and education). Upon finishing, participants were debriefed, and the study was concluded. All participants were entered into the €50 raffle, regardless of the accuracy of their veracity judgements. Participation in the study took approximately 30 minutes.

To evaluate the accuracy of participants' veracity judgements, the binary truth–lie judgements were transformed into incorrect and correct veracity decisions, according to the ground truth.

Results

Motivation, experimental realism and self-perceived lie detection ability

Several 2 (baseline instruction: present, absent) × 2 (veracity of the critical element: truth, lie) analyses of variance (ANOVAs) were conducted on participants' responses to a series of 7-point Likert scales (1 = *not at all*, 7 = *very*). These analyses revealed that participants, on average, were highly motivated ($M = 6.29$, $SD = 0.87$), with no significant differences between baseline and veracity conditions, $F(3, 144) = 1.07$, $p = .366$, $\eta_P^2 = .022$. On average, participants reported to answer questions honestly ($M = 6.82$, $SD = 0.45$), with no significant differences between baseline and veracity conditions, $F(3, 144) = 0.29$, $p = .830$, $\eta_P^2 = .006$. On average, participants found the instructions very clear ($M = 6.49$, $SD = 0.80$), with no significant differences between baseline and veracity conditions, $F(3, 144) = 1.26$, $p = .290$, $\eta_P^2 = .026$. We also observed that participants, on average, found the statements realistic ($M = 5.39$, $SD = 1.24$), with no significant differences between baseline and veracity conditions, $F(3, 144) = 0.35$, $p = .789$, $\eta_P^2 = .007$. Additionally, participants self-reported to be average lie detectors ($M = 4.10$, $SD = 1.14$), with no significant differences between baseline and veracity conditions, $F(3, 144) = 1.11$, $p = .345$, $\eta_P^2 = .023$.

Veracity judgements

Participants' veracity judgements were analysed using a two-way between-subjects ANOVA on the Likert scale judgements, and a logistic regression on the binary data. In addition, we also subjected the binary judgements to a two-way between-subjects ANOVA. Although duplicate, we report this analysis to allow for a comparison with previous research and because it allows for the calculation of Bayes factors, which helps with the interpretation of the results. The Bayesian factors (*BF*; for interpretation, see Jarosz & Wiley, 2014; Lee & Wagenmakers, 2013) are reported in line with the cut-off points outlined by Jeffreys (1961). The approximate evidence

categories are as follows: values between 1 and 3 indicate weak evidence for the alternate or null hypothesis, between 3 and 10 indicate strong evidence, and above 10 are considered very strong evidence. The interaction model within JASP combines both main effects and the interaction effect; therefore, evidence for the interaction term individually was calculated by dividing the interaction model by the main factors (e.g. Wagenmakers et al., 2016). For ease of interpretation, BF_{10} is used to indicate the Bayes factor as evidence in favour of the alternative hypothesis, whereas BF_{01} is used to indicate the Bayes factor as evidence in favour of the null hypothesis.

Table 1 displays the means and standard deviations of both the binary and Likert judgements. Overall, the accuracy of participants' binary judgements, where 0 represents incorrect and 1 represents correct judgements, did not differ significantly from chance level ($M = .53$, $SD = .50$), $t(147) = 0.82$, $p = .413$, $d = 0.06$. To examine whether the baseline instruction increased participants' ability to accurately discriminate between lies and truths, we conducted a 2 (baseline instruction: present, absent) × 2 (veracity of the critical element: truth, lie) between-subjects ANOVA on the accuracy of participants' binary veracity judgements. Contrary to our hypothesis, the main effect of the baseline instruction was not significant, $F(1, 144) = 0.26$, $p = .613$, $\eta_P^2 = .002$; $BF_{01} = 5.06$, meaning that participants who received the baseline-present instructions ($M = .55$, $SD = .50$, 95% confidence interval, CI [.44, .67]) were not significantly more accurate in their veracity judgements than participants who received the baseline-absent instructions ($M = .51$, $SD = .50$, 95% CI [.40, .63]). This analysis revealed a main effect of the veracity of the critical element, $F(1, 144) = 10.33$, $p = .002$, $\eta_P^2 = .067$; $BF_{10} = 18.49$, with lies ($M = .66$, $SD = .48$, 95% CI [.55, .77]) being judged more accurately than truths ($M = .41$, $SD = .49$, 95% CI [.29, .52]). Finally, the veracity of the critical element by baseline instruction interaction effect was also not significant, $F(1, 144) = 1.40$, $p = .238$, $\eta_P^2 = .010$; $BF_{01} = 2.32$, indicating that the baseline instruction had no differential effect on the accuracy of participants' veracity judgements for lies and truths.

The logistic regression on the binary judgement with predictor variables (baseline instruction: present, absent, and veracity of the critical element: truth, lie) and the accuracy of participants' binary veracity judgements as the dependent measure, revealed a significant overall model, $\chi^2(2) = 10.18$, $p = .006$, Nagelkerke $R^2 = .09$. Contrary to our hypothesis, the baseline instruction was not a significant predictor of classification accuracy ($p = .609$, 95% CI [0.43, 1.64]). The veracity of the critical element, however, was a statistically significant predictor ($p = .002$): participants who evaluated a statement containing a deceptive critical element had 2.88 times higher odds (95% CI [1.47, 5.63]) of making a correct veracity judgement than those who evaluated a statement containing a truthful critical element.

The 2 (baseline instruction: present, absent) × 2 (veracity of the critical element: truth, lie) between-subjects ANOVA on participants' Likert scale veracity judgements revealed no significant differences. We did not observe a significant main effect of the baseline instruction, $F(1, 144) = 0.95$, $p = .332$, $\eta_P^2 = .007$; $BF_{01} = 3.69$ ($M_{\text{baseline-present}} = 4.27$, $SD = 1.54$, 95% CI [3.92, 4.62] versus $M_{\text{baseline-absent}} = 4.03$, $SD = 1.53$, 95% CI [3.68, 4.38]). Nor did we find a significant main effect of the veracity of the critical element, $F(1, 144) = 0.75$, $p = .388$, $\eta_P^2 = .005$; $BF_{01} = 4.04$ ($M_{lies} = 4.26$, $SD = 1.50$, 95% CI [3.91, 4.61] versus $M_{truths} = 4.04$, $SD = 1.57$, 95% CI [3.69, 4.39]). Finally, the Veracity × Baseline interaction effect was not significant, $F(1, 144) = 3.80$, $p = .053$, $\eta_P^2 = .026$; $BF_{01} = 0.84$.

Discussion

It is well documented that observed lie detection rates hover around 50% (e.g. Bond &

DePaulo, 2006). We replicated this finding. We found that when making a binary judgement, participants were significantly better at detecting lies than at detecting truths. Contrary to our hypothesis, however, we found that participants who were instructed to utilize a within-statement comparable baseline did not outperform the control group in terms of overall lie–truth discrimination accuracy.

Previous research revealed that the comparable truth baseline technique can enhance observers' judgement accuracy (Caso, Palena, Vrij, et al., 2019). A possible explanation for the divergence in findings between this study and our own stems from how the statements were generated. In the experiment by Caso, Palena, Vrij, et al. (2019), interviewees reported about experimental tasks they had completed. Our study, in contrast, used statements about participants' experienced activities on a particular day. Consequently, our participants were mostly unconstrained in their reports. These paradigms differ systematically in the source of deception, which was either scripted by the researcher (Caso, Palena, Vrij, et al., 2019, i.e. a scripted task; Vrij, 2008) or drawn freely from the participant's own experience (current study, i.e. an autobiographical task; Sporer & Sharman, 2006). When lies are self-generated, the deceiver can elaborate with personal experience, whereas lies designed by the researcher cannot be so easily embellished. As a consequence, it is possible that the statements resulting from the different sources of deception were perceived differently by the lie-detectors. Scripted tasks, relative to autobiographical tasks, may have been more straightforward to evaluate, which perhaps contributed to the incongruent findings between the present study and that of Caso, Palena, Vrij, et al. (2019). A second explanation for these results could be derived from our ground truth manipulation check. Compared to interviewees who reported entirely truthful recollections of their day, interviewees who embedded a lie reported lower ratings of truthfulness for the general, truthful portion of their statement. Thus, the truthful baseline of liars' statements may have been more comparable with their lies, which could have weakened a lie-detector's ability to make accurate decisions. Although this may accurately reflect real-world conditions in which interviewees interweave truths and lies (e.g. Vrij, 2008), for the purposes for testing the efficacy of within-statement baselining future research should utilize a more controlled method for establishing baseline truthfulness.

The analysis of variance of the binary veracity judgements allows for a comparison to the results of previous work (Caso, Palena, Vrij, et al., 2019). These authors found that using a comparable truth baseline led to significantly above-chance levels in total accuracy for distinguishing truth-tellers from liars ($d = 0.34$). We found no effect of the baseline instruction with a Bayes factor indicating substantial evidence for the absence of an effect. To directly compare to previous research, we also calculated a Cohen's d effect size, which reaffirmed a very small effect ($d = 0.08$) of the baseline instruction on participants' binary judgements. The finding from the Likert judgements are, however, more ambiguous. This measure also indicated no effect of the baseline instruction, but the veracity by baseline interaction revealed a p value of .053 and was accompanied by an inconclusive Bayes factor of 0.84 in favour of the null hypothesis. This pattern suggests that although we did not find evidence that the within-statement baseline comparison was an effective lie detection tool, we also cannot rule out that it may have an effect that was too small for our study to pick up. Future research should replicate our study with an increased sample size. Although our power analysis indicated that we had sufficient power to detect an effect size of similar nature of Caso, Palena, Vrij, et al. (2019) had it been present, we may have been underpowered to detect a smaller effect size.

Current methods of verbal baselining may require further refinement before they can be reliably used to detect deception. In two

studies to date (the current study and Caso, Palena, Carlessi, et al., 2019), instructing observers to make comparisons of the verbal content between known-truthful and target portions of a statement did not improve the accuracy of credibility assessments. In both experiments, participants in the baseline conditions were invited to search for deviations or patterns that might be indicative of deceit. Alternatively, explicitly training participants to look for empirically valid verbal deception cues, such as the richness of detail (e.g. Luke, 2019) or the verifiability of information (e.g. Nahari, Vrij, & Fisher, 2014), may increase the effectiveness of baselining.

In sum, we did not find evidence that the within-statement baseline technique can enhance deception detection accuracy. Additional work is required to refine the technique and determine its true efficacy across different contexts.

Author note

This research is supported by a fellowship awarded from the Erasmus Mundus Joint Doctorate Program of The House of Legal Psychology (EMJD-LP) with Framework Partnership Agreement (FPA) 2013-0036 and Specific Grant Agreement (SGA) 2016-1339 to Brianna L. Verigin.

Ethical standards

Declaration of conflicts of interest

Brianna L. Verigin has declared no conflicts of interest

Ewout H. Meijer has declared no conflicts of interest

Aldert Vrij has declared no conflicts of interest

Ethical approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee [173_01_11_2016_S4] and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent

Informed consent was obtained from all individual participants included in the study.

Supplemental material

Supplemental material is available via the "Supplementary" tab on the article's online page (https://doi.org/10.1080/13218719.2020.1767712).

Note

1. The interviews included in the present study were a mean length of 4 minutes and 28 seconds ($SD = 2.50$; range: 1 minute and 50 seconds to 15 minutes and 22 seconds).

ORCID

Brianna L. Verigin http://orcid.org/0000-0001-8941-8398

Ewout H. Meijer http://orcid.org/0000-0001-9590-3699

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

References

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

Bond, C.F., & DePaulo, B.M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc*, *10*(3), 214–234. doi:10.1207/s15327957pspr1003_2

Brandt, D.R., Miller, G.R., & Hocking, J.E. (1980a). Effects of self-monitoring and familiarity on deception detection. *Communication Quarterly*, *28*(3), 3–10. doi:10.1080/01463378009369370

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

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

Caso, L., Palena, N., Carlessi, E., & Vrij, A. (2019). Police accuracy in truth/lie detection when judging baseline interviews. *Psychiatry, Psychology and Law, 1–10*, *26*(6), 841–850. doi:10.1080/13218719.2019.1642258

Caso, L., Palena, N., Vrij, A., & Gnisci, A. (2019). Observers' performance at evaluating truthfulness when provided with comparable truth or small talk baselines. *Psychiatry, Psychology, and Law: An Interdisciplinary Journal of the Australian and New Zealand Association of Psychiatry, Psychology and Law*, *26*(4), 571–579. doi:10.1080/13218719.2018.1553471

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: L. Erlbaum Associates.

Comadena, M.E. (1982). Accuracy in detecting deception: Intimate and friendship relationships. In M. Burgoon (Ed.), *Communication yearbook* 6 (pp. 446–472). Beverly Hills, CA: Sage.

DePaulo, B.M., Lindsay, J.J., Malone, B.E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin*, *129*(1), 74–112. doi:10.1037//0033-2909.129.1.74

Ekman, P., & Friesen, W.V. (1974). Detecting deception from the body or face. *Journal of Personality and Social Psychology*, *29*(3), 288–298. doi:10.1037/h0036006

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

Faigman, D.L., Monahan, J., & Slobogin, C. (2014). Group to individual (G2i) inference in scientific expert testimony. *The University of Chicago Law Review*, 417–480.

Feeley, T.H., deTurck, M.A., & Young, M.J. (1995). Baseline familiarity in lie detection. *Communication Research Reports*, *12*(2), 160–169. doi:10.1080/08824099509362052

Fisher, A.J., Medaglia, J.D., & Jeronimus, B.F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(27), E6106–E6115. doi:10.1073/pnas.1711978115

Frank, M.G., Yarbrough, J.D., & Ekman, P. (2006). Investigative interviewing and the detection of deception. In T. Williamson (Ed.), *Investigative interviewing: Rights, research and regulation* (pp. 229–255). Devon, United Kingdom: Willan Publishing.

Hayano, D.M. (1980). Communicative competency among poker players. *Journal of Communication*, *30*(2), 113–120. doi:10.1111/j.1460-2466.1980.tb01973.x

Inbau, F.E., Reid, J.E., Buckley, J.P., & Jayne, B.C. (2013). *Criminal interrogation and confessions* (5th ed.). Burlington, MA: Jones & Bartlett Learning.

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*(1), 2–9. doi:10.7771/1932-6246.1167

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford: Oxford University Press.

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

Luke, T.J. (2019). Lessons from pinocchio: Cues to deception may be highly exaggerated. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *14*(4), 646–671. doi:10.1177/1745691619838258

McCornack, S.A., & Parks, M.R. (1986). Deception detection and relational development: The other side of trust. In M. L. McLaughlin (Ed.), *Communication yearbook* (Vol. 9, pp. 377–389). Beverly Hills, CA: Sage. doi:10.1080/23808985.1986.11678616

Nahari, G., Vrij, A., & Fisher, R.P. (2014). The verifiability approach: Countermeasures facilitate its ability to discriminate between truths and lies, countermeasures facilitate its ability to discriminate between truths and lies. *Applied Cognitive Psychology*, *28*(1), 122–128. doi:10.1002/acp.2974

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

Russano, M.B., Narchet, F.M., Kleinman, S.M., & Meissner, C.A. (2014). Structured interviews of experienced HUMINT interviewers. *Applied Cognitive Psychology*, *28*(6), 847–859. doi:10.1002/acp.3069

Sporer, S.L., & Sharman, S.J. (2006). Should I believe this? Reality monitoring of accounts of self-experienced and invented recent and distant autobiographical events. *Applied Cognitive Psychology*, *20*(6), 837–854. doi: 10.1002/acp.1234

U.S. Department of the Army. (2006). *Field manual 2–22.3 (FM 34–52) Human intelligence collector operations*. Washington, DC: Headquarters, Department of the Army.

Verigin, B. L., Meijer, E. H., & Vrij, A. (2020). Embedding lies into truthful stories does not affect their quality. *Applied Cognitive Psychology*, *34*, 516–525. doi: 10.1002/acp.3642

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

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

Wagenmakers, E.J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., … Morey, R.D. (2016). Bayesian inference for psychology. Part II: Example applications with JASP, 1–26. Retrieved from http://maarten-marsman.com/wpcontent/uploads/2017/04/WagenmakersEtAlPartII.pdf

Appendix
Experimental instructions

Dear Participant,

Please imagine yourself in the role of a district detective in the Criminal Investigation Department of the Limburg Police. You are the lead detective for an investigation into a violent burglary that occurred approximately one week ago. Your colleague just finished interviewing a suspect, and your job is to review the interview statement and to assess the suspect's credibility. The interviewee has provided an alibi statement for the entire day in question, from morning to evening, but the critical period of time (i.e. the "statement of interest") that you are most interested in is from ***1:00pm to 3:00pm****.*

Your task is to read the statement carefully and to make a decision about the truthfulness of the critical statement of interest (i.e. the highlighted information, the period of time from 1:00pm to 3:00pm). You will then respond to several questions regarding your decision.

It is extremely important that your decision is correct, if not, either the perpetrator gets away with the crime OR you may send an innocent person to jail. Plus, if you make the correct decisions regarding truthfulness, you will be entered into a raffle to win €50.

Additional instruction for baseline-present condition:

> *As the lead investigator, you have access to other sources of information for this case. This evidence confirms that* ***the "general" alibi statement before 1:00pm and after 3:00pm is truthful.*** *Please use this knowledge to compare the "general" part of the interviewees' alibi to the "statement of interest from 1:00pm to 3:00pm." Try to identify any patterns or changes in the verbal content between the general alibi and the statement of interest that may indicate how credible the suspect's story from 1:00pm to 3:00pm is.*