A re-analysis that supports our main results: A reply to Levine et al.

Aldert Vrij

Hartmut Blank

Ronald P. Fisher

Author Note

Aldert Vrij and Hartmut Blank, Department of Psychology, University of Portsmouth; Ronald Fisher, Department of Psychology, Florida International University.

Correspondence concerning this article should be addressed to Aldert Vrij, Department of Psychology, University of Portsmouth, King Henry Building, King Henry 1 Street, PO1 2DY, Hants, United Kingdom. Email aldert.vrij@port.ac.uk

Abstract

Levine et al. (2017) criticized our meta-analysis, but their conclusion was the same as ours: The cognitive approach to lie detection results in a modest improvement. We address and dismiss Levine et al.'s (2017) three criticisms. Regarding the 'confound,' in our meta-analysis we averaged the results of two cells on statistical grounds, which does not constitute a confound in statistical terms. Regarding 'aberrant controls,' that depends entirely on the benchmarks selected and type of statistical test and meta-analysis used. Regarding 'unreliable data,' the claim that there is a positive relationship between 'unreliable' data and total accuracy in the cognitive lie detection conditions is not even supported by their own data (p = .16). We conclude with a request to Levine et al. to focus on our shared aim: To develop interview protocols that enable lie detection.

Key words: cognitive lie detection; meta-analysis; Interviewing to detect deception

A re-analysis that supports our main results: A reply to Levine et al.

Levine et al. (2017) came to the same main conclusions as we did: The cognitive approach to lie detection results in a modest improvement (*d* = around .40) in the ability to distinguish truth tellers from liars, and the same improvement is made when the veracity assessments are made by human judges or through statistical algorithms based on objective criteria.[i] We therefore consider the points they raised primarily cosmetic. For example, we averaged the human judges and statistical algorithms accuracy rates, resulting in 56% (control condition) and 71% (cognitive lie detection condition) accuracy rates. Levine et al. (2017) argued we should never have averaged the scores. Our meta-analysis did not focus on differences in accuracy rates obtained by human judges or through statistical algorithms; it focussed on the improvement cognitive lie detection makes on veracity judgements. Since the improvement is similar for human judges and statistical algorithms, presenting these averaged scores is reasonable.

'Confounded' dependent variables, 'aberrant controls' and 'unreliable' data

In their abstract Levine et al. (2017) claimed that '*Vrij et al.'s analyses confounded dependent variables, capitalized on aberrant controls and used unreliable data to inflate support*.' There are problems with these three claims. In our meta-analysis, we averaged the results of two experimental cells on statistical grounds, which does not constitute a confounding variable in statistical terms.

The 'aberrant' controls depend on the types of statistical analysis and meta-analysis used (see Appendix 1) and on the benchmark selected. Levine et al. chose a 73% benchmark based on personal communication with Bond rather than the accuracy rate benchmark cited in Hartwig and Bond (2014) (67.68%). The accuracy rates in the control conditions do not differ statistically from the 68% benchmark (see

Appendix 1). Deciding to report benchmarks not published in the literature opens up a debate as to which benchmark to use, which we do not consider useful.

The more important question is what low accuracy rates in control conditions actually mean. It tells us that discriminating between truth tellers and liars is difficult in the situations examined in those studies. In some situations, lie detection is easier than in others. Levine et al. (2014) reported deception accuracy rates up to 100%. We have argued that in their deception scenario, truth tellers can demonstrate relatively easily that they are telling the truth, whereas liars are faced with a difficult situation to convince the interviewer that they are not lying (Vrij, Meissner, & Kassin, 2015). In such situations, professional investigators should have little difficulty in distinguishing truth tellers from liars when interviewing them, as Levine et al. (2014) found. Professional investigators tell us that they are not interested in "easy scenarios'" because they already possess the skills to resolve those scenarios. They are interested in difficult scenarios in which truth tellers find it difficult to demonstrate that they are telling the truth and liars are capable to give honest-sounding answers. The control data in the cognitive lie detection studies show that researchers successfully created and tested such scenarios. If interview techniques in such difficult scenarios lead to a modest improvement in accuracy rates and raise these accuracy rates to levels well above chance (Levine et al., 2017; Vrij, Fisher, & Blank, 2017), we consider this to be a success.

Since different scenarios result in different accuracy rates, we find it more valuable to compare the results of the experimental conditions with the results of the control conditions in the same experiment than to compare the results of experimental conditions with benchmark accuracy rates obtained from different experiments found

in the wider literature, and especially when different researchers use different benchmark rates.

Levine et al. (2017) expressed concern about the negative relationship between accuracy rates in the cognitive lie detection conditions and the number of judgments made in these conditions; however, their claim was not even supported by their own analysis (p = .16). The weakest finding for a cognitive lie detection condition (54%) was found in the study with the most judgments (N = 864, Vrij, Mann, Leal, & Fisher (2010), for which we have a theoretical explanation. The 'instructing interviewees to look into the eyes of the interviewer' manipulation was not very strong, as it is virtually impossible for interviewees to look the interviewer into the eyes all the time when talking (Kajimura & Nomura, 2016). We therefore left this technique out of a cognitive lie detection training (Vrij, Leal, Mann, Vernham, & Brankaert, 2015). The negative relationship between accuracy rates in the cognitive lie detection conditions and the number of judgments made in these conditions was not even close to being significant (p = .52) when Vrij, Mann, Leal, and Fisher (2010) was excluded.

Three irrelevant 'concerns'

In the re-examination part of their reply Levine et al. raised three irrelevant concerns. The standard interview protocols varied, in part, because the deception scenarios in these studies varied and different scenarios sometimes require different interview protocols. The interview techniques in the experimental conditions also varied and we have always stated that the cognitive approach to lie detection should include different techniques. Regarding the 1,500 to 3,000 judgments required in a lie detection study (second 'concern'), ideally each observer assesses only one statement because only that reflects the real-life situation in which an observer has to make a

decision in an individual case. This would mean 1,500-3,000 participants, which is unrealistic. The cross-validation 'concern' does not apply to the primary analyses (comparing experimental conditions with control conditions), because the lack of cross-validation will affect the control and experimental conditions in a single experiment in similar ways.

Final Thought

Whereas we welcome critical evaluation of our work, and we think it healthy for the scientific community, not all criticism is valid, such as Levine et al.'s (2017) analysis. Given the importance of developing new theory-based, empirically supported approaches to detecting detection in the modern, security-conscious world (Vrij, Meissner et al., 2017), we hope that in the future Levine and colleagues will focus on developing theory-based interview protocols that facilitate lie detection. After all, Levine and colleagues have in common with us that we believe that lie detection can be improved through adequate questioning (Levine et al., 2014).

References

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

Colwell, K. James-Kangal, N., Hiscock-Anisman, C., & Phelan, V. (2015). Should police use ACID? Training and credibility assessment using transcripts versus recordings. *Journal of Forensic Psychology Practice, 15*, 226-247. Doi: 10/1080/15228932.2015.1035187

Hartwig, M., & Bond, C. F. (2014). Lie detection from multiple cues: A meta-analysis. *Applied Cognitive Psychology*, *28*, 661-667. DOI: 10.1002/acp.3052.

Hays, W. L. (1994). *Statistics* (5th ed.). Belmont, CA: Wadsworth.

Kajimura, S., & Nimura, M. (2016). When we cannot speak: Eye contact disrupts resources available to cognitive control processes during verb generation. *Cognition, 157*, 352-357. Doi: 10.106/j.cognition,2016.10.002

Levine, T. R. (2014). Active deception detection. *Policy Insights from Behavioral and Brain Sciences, 1,* 122-128. Doi: 10.1177/2372732214548863

Levine, T. R., Clare, D. D., Blair, J. P., McCornack, S., Morrison, K., Park, H. S. (2014). Expertise in deception detection involves actively prompting diagnostic information rather than passive behavioral observation. *Human Communication Research, 40*, 442-462. doi:10.1111/hcre.1203

Ormerod, T. C., & Dando, C. J. (2014). Finding a needle in a haystack: Toward a psychologically informed method for aviation security screening. *Journal of Experimental Psychology*: *General,* 144, 76–84. Doi:10.1037/xge0000030

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

Vrij, A., Leal, S., Mann, S., Vernham, Z., & Brankaert, F. (2015). Translating theory into practice: Evaluating a cognitive lie detection training workshop. *Journal of Applied Research in Memory and Cognition, 4*, 110-120. doi:10.1016/j.jarmac.2015.02.002

Vrij, A., Mann, S., Leal, S., & Fisher, R. (2010). "Look Into My Eyes": Can an instruction to maintain eye contact facilitate lie detection? *Psychology, Crime, & Law*, *16*, 327-348. DOI 10.1080/10683160902776843

Vrij, A., Meissner, C. A, Fisher, R. P., Kassin, S. M., Morgan III, A., & Kleinman, S. (2017). Psychological perspectives on interrogation. *Perspectives on Psychological Science*. DOI 10.1177/1745691617706515

Vrij, A., Meissner, C., A. & Kassin, S. M. (2015). Problems in expert deception detection and the risk of false confessions: No proof to the contrary in Levine et al. (2014). *Psychology, Crime, & Law*, *21*, 901-909. DOI 10.1080/1068316X.2015.1054389.

Appendix 1: Type of statistical test and meta-analysis

Which statistical test to use?

Levine et al.'s (2017) 'aberrant controls' claim is based on one-sample *t*-tests (and related confidence intervals) of control performance against the Bond and DePaulo (2006) 54% benchmark (and later similar tests against other benchmarks). The 1,527 *df* test at the judgement level aggregates both independent and dependent (i.e. within-subjects) data, which is usually considered problematic from a statistical point of view. Also, a general problem with using *t*-tests is that they are for normally distributed data; however, accuracy data are not normally distributed, as they can take on only the values 0 and 1. We performed analyses using binomial testing, an adequate approach for data based on categorical decisions. Specifically, we conducted binomial tests against relevant benchmarks (see e.g. Hays, 1994, p. 259) on the basis of the total sample size (thus avoiding the dependency problem associated with repeated measurement within participants).

For human judges, it turned out that our 48% standard approach total accuracy does not differ significantly from the benchmark (54%, Bond & DePaulo, 2006), $z = 1.94$, $p = .052$, but the cognitive approach total accuracy (62%) does, $z = 2.83$, $p = .005$. For computer classification, our 64% standard accuracy does not differ significantly from the benchmark (68%, Hartwig & Bond, 2014) ($z = 1.51$, $p = .13$), but the cognitive approach total accuracy (79%) does, $z = 5.06$, $p < .001$. Finally, when we use the 73% benchmark used by Levine et al. (2017), the total accuracy rate in the standard approach control condition is significantly lower than the benchmark ($z = 3.31$, $p = .001$), but the cognitive approach total accuracy rate is still significantly higher than this 73% benchmark ($z = 2.85$, $p = .004$).

Which meta-analytic approach to use?

Some of Levine et al. (2017)'s points are conditional on their exclusive focus on Cohen's *d* as a measure of effect size. For illustration, Levine et al. (2017) focused on *d*-scores when claiming that the standard-cognitive difference is the same size as the 50% to 54% difference in Bond & De Paulo (2006). This claim strongly depends on using *d* as the effect size; the standard-cognitive difference is much larger (1.97 vs. 1.17), and also more in line with the (weighted) accuracy percentage differences, when using odds ratios as the effect size. In our paper we discussed at length why odds ratios are an appropriate effect size measure (see footnote 1).

---

[i] The same improvement is made in both types of judgement, which does not support Levine et al. 's claim that the effects should be larger for statistical algorithms than for human judges. Levine et al.'s claim is incorrect because it does not take into account that these statistical algorithms are often based on very few cues (sometimes on only one cue, such as amount of detail). The gain someone could get from analyzing a single cue is limited. If human judges based their judgements on more than one cue (for example not only on the amount of detail but also on something not objectively coded in a study, such as the plausibility of such details) and if the cognitive approach enhanced the discriminatory power of both these cues, human judges are likely to profit more from the cognitive approach than a single cue objective criteria algorithm could ever do.

Also, their argument in endnote 3 is misleading as the objective criteria algorithms used in four of the seven studies were based on just one cue and in one study on two cues. The high mean (3.28) was driven by one study in which 13 cues

were entered (in the seventh study four cues were entered). A similar reasoning applies to the number of cues examined in the studies.