Commentary

Accurate expert deception detection: Faulty premises in Vrij et al. (2015)

Timothy R. Levine 1* and J. Pete Blair 2

1 Department of Communication Studies, University of Alabama at Birmingham, United States
2 Department of Criminal Justice, Texas State University, United States

Received 24 July, 2017; Accepted 4 January, 2018

This essay is a rejoinder to a commentary by Vrij et al. (2015) that was critical of Levine et al.’s (2014) experiments showing exceptionally high deception detection accuracy by expert interrogators. We contend that highly accurate deception detection is possible under conditions where contextualized communication content is diagnostic and where an interrogator is allowed to actively solicit honest confessions. Such conditions are often met outside the lab. Vrij et al. dismiss our findings as methodological artifact. They argue that we did not demonstrate expertise because the task was so easy that anyone could perform to perfection. This essay shows that the premises of Vrij et al.’s arguments are false and their arguments falter as a result. High accuracy is possible with skilled expert interrogation.

Key words: Lies, deception, confessions.

INTRODUCTION

We (Levine et al., 2014) recently reported results of two experiments documenting exceptionally high levels of accuracy by expert interrogators under conditions where honest confession-seeking and communication content were viable approaches to assessing the guilt and innocence of interviewees. Because our results were much stronger than previous findings, we anticipated skepticism. Until our findings are replicated by other researchers and labs, caution is scientifically warranted. But, while skepticism is understandable and scientific debate is both healthy and welcome, scientific arguments need to be logically sound. To be sound, the premises of the argument must be true and the conclusions must follow from the premises. Vrij et al. (2015)’s arguments contending that our conclusions lack scientific contribution and are dangerous are based on a series of false premises. Here we show that their critique is a string of fallacious straw person arguments.

We agree with Vrij et al. (2015) that research prior to 2006 documents that accuracy in deception detection experiments is approximately 54% plus or minus 10%, and that past research has found that experts perform similarly to students. This was explicit in the literature review of our 2014 paper. We were clear in our original article that the preponderance of prior evidence predicted poor accuracy (but see Levine, 2015 for summaries of
more recent findings).

We argue, however, that the large body of prior findings summarized in meta-analysis applies specifically to “cue-biased” lie detection. We theorize (Levine, 2014) that cue-based lie detection invariably produces slightly-better-than-chance accuracy because cues are not diagnostic for most senders, cues are not constant across messages, senders, and contexts, and cue findings do not replicate across studies.

Fortunately, cues are not the only way to detect lies. Outside the lab, most lies are detected either through the use of evidence or through honest confessions (Park et al., 2002). A recent replication of Park et al. with police shows that this is how law enforcement officers detect most lies too (Masip and Herrero, 2015). Thus, we reasoned, if deception detection expertise exists, it is not based on either the presence of cues in senders or skilled reading of cues by experts. Instead expertise rests on prompting diagnostic, contextualized, communication content - including honest admissions. Expert interrogators can be really good at persuading people to be honest and short of that, they make it difficult for interviewees who are actually being uncooperative to feign cooperation and honesty. They do this through skilled question asking and persuasive communication that cannot be captured in pre-scripted questions. It is a truism in communication that effectiveness requires adaptation to the person receiving the message. Scripting, of course, is antithetical to this. The ability to go off script is therefore critical.

We (Levine et al., 2014) conducted two studies using our cheating experiment in which subjects were provided the opportunity to cheat on a trivia game for a cash incentive. Potential cheaters were then questioned by experienced interrogators who attempted to correctly ascertain the truth. The experts were highly adept at ascertaining actual guilt or innocence. Accuracy was over 95% in each of two experiments. Undergraduate students watching the expert interviews were also able to correctly ascertain guilt or innocence with impressive accuracy (between about 80 and 93% depending on the specific set of videos).

We understand that our results were unusually strong and seem inconsistent with prior research. We also realize that (1) obtaining such high accuracy with an expert trained in the Reid Technique (our first experiment) and (2) our use of accusatory questioning (including false evidence) which did not yield false confessions, seem to contradict the consensus view in legal and criminal psychology. On face, our findings appear to contradict the findings and claims of our critics’ programs of research. Our critics (Vrij et al., 2015) would have readers believe that their own findings are unassailable and our findings are pure artifact easily dismissed as unscientific.

The question is, was the high accuracy in our experiments wholly attributable to the idiosyncrasies of our research paradigm as Vrij et al. (2015) suggest or was it because the experts linguistically created an unusually diagnostic environment though skilled questioning? If Vrij or one of his colleagues had done the questioning, could they have done as well as any of our experts? We think it is the expert’s skill at question-asking that made the accuracy so high. Vrij et al. contend that it is not the experts or what they did but instead the nature of our method for soliciting truths and lies that produced the strong results. According to them, anyone could do well in our method by just asking two simple questions (how many answers did the subject get right and how did the subject know the answers). We think, the very fact that they assert this suggests that they would not perform nearly as well as our experts. Let us explain why.

A bestiary of straw beings

Vrij et al.’s (2015) key argument comes down to this: “The problem with Levine et al.’s paradigm is that it is easy for truth tellers to demonstrate their innocence but virtually impossible for liars to tell a convincing lie” (p. 12). This argument, however, reflects a series of interrelated misconceptions and misrepresentations of our research. Vrij et al.’s premises are either false or exaggerations and the inaccuracies are easy to demonstrate. Their arguments against our findings are straw person arguments. They criticize flaws that didn’t exist in our method and that cannot explain our results.

In order for Vrij et al.’s concerns about our work to be sound, a series of three critical conjectures must be accepted; otherwise their argument unravels. Their argument requires:

1) Honest non-cheaters knew and could convincingly explain how they knew the correct answers (that is, honesty was unusually easy).
2) Liars could not explain how they got answers correct (that is, lying was too hard).
3) Honest interviewees could prove their innocence (or thought they could), so they did not falsely confess.

It was definitely true that some liars could not create plausible explanations and that many honest interviewees could remember how they learned the facts in question. But, this was far from universal. First, the honest non-cheaters frequently could not recall how they learned a fact and frequently (perhaps even more so than liars) answered honestly “I don’t know”, “I just knew it” or “I just guessed.” We ask the reader to think of all the random trivia in their own memory. What proportion could you, on the spot, under the most intense questioning you have ever faced in your life, convincingly recall and explain how it is that you came to know the fact? It is just not that easy for the honest interviewees. It should not be news to readers that memory is imperfect. Readers
should be highly skeptical of arguments premised on the idea that unimportant information learned in the distant past is easy to recall and communicate convincingly. Easy recall and communication of the sources of trivial knowledge is implausible on its face. Watching our videos shows that the task was not easy for many honest interviewees.

When honest subjects did know how they knew some fact, the reasons were often quite similar to the reasons provided by many cheating liars. "I learned it in class" or "I saw it on the Discovery Channel" were frequent answers for both cheaters and non-cheaters alike. It is not that hard for a cheater to explain that she learned the answer to a history question in a history class. "I remembered it because it was on the test" is not a hard lie. Again, watching our videos shows that Vrij et al.'s (2015) depiction is not what actually happened.

Regarding honest interviewees being able to prove their innocence, there was no way for an innocent student to prove their innocence to the expert. Vrij et al.'s (2015) assertion to the contrary is flat out false. There was only one witness, their partner, and they were not in the room when the participant was being questioned. Honest participants and lying cheaters alike claimed that their partner would exonerate them. And, in our first study, a false evidence ploy was used by the interrogator in every interview. Honest and guilty subjects alike were told (falsely) that their partner, the only person who could collaborate their story, had implicated them. The idea that exculpatory evidence was at hand is wrong. The reverse was actually the case. Anyone who would take the time to watch our videos can see that Vrij et al.'s (2015) conjectures are objectively false occurring only in their imagination.

The reason we didn’t get false confessions is, at least in part, that interviewers made it clear that they were interested in the truth; they wanted confessions only if there was guilt to confess. We did have a couple non-cheating subjects offer to confess, but when asked by an expert if the offered confession would be the truth, all false confession offers were quickly retracted by the non-cheating subjects.

Clearly, the heart of Vrij et al.'s critique is a set of straw people. Honest participants often could not convincingly explain how they knew the answers. Liars often had explanations for how they knew the answers that were quite similar to those given by truth-tellers. Honest interviewees could not prove their innocence (particularly in experiment one). Our experiments were videotaped and watching the tapes disproves our critic's arguments. Because Vrij et al.'s conclusion rest on false premises, their argument is fallacious.

Outside of the core of the Vrij et al. (2015) critique there are also a number of additional ancillary straw people presented to bolster their argument. The first is that a simple decision rule about the number correct can yield 80% accuracy. Going by the number correct in the trivia game was not nearly as simple as Vrij et al. (2015) make it sound. Just like Ariely's (2013) subjects, many of our cheaters did not cheat that much; only a question or two. Using the number correct is clearly diagnostic, but even if a simple decision rule yields 80% accuracy, this still cannot explain the accuracy observed in either experiment. For study 1, the probability of the observed success with a 0.8 prior probability is $p < 0.0006$. The probability of correctly guessing 87 out of 89 interviews correct (97.8%) in study 2 is $p < 0.0000006$. It should also be noted that some of our experts were 100% correct. As such, they faced a ceiling effect and the probabilities stated above could overstate the probability of achieving that level of accuracy. Thus, statistically Vrij et al.'s simple decision rule cannot explain our findings. If that is their strategy, then across our two studies, the smart money would be on our experts, not Vrij et al. at odds of a million-to-one.

Even more importantly, the Vrij et al. (2015) critique is offered with the benefit of hindsight not available to our experts. There was no way to know at the time of the interviews just how diagnostic the number correct would be. In our past experiments with the trivia game using the same task, non-cheaters have scored as high as 6 out of 10, and cheaters have obtained zero correct (which happened once in Experiment 2). With hindsight, it is known that using a number correct heuristic yields 80%, but successful experts could not rely on that alone. It would guarantee a long-run error rate of 20%. Further, interviewers were not told how many the questions the participant answered correctly. Thus, the interviewer could not know that subjects were not lying about how many they got right, which questions were correct, or who got them right, their partner or themselves.

Equally importantly, our experts did not know what the cheating base-rate would turn out to be. This too was only known after the data were collected and analyzed. The cheating base rate in the first study (~12%) was substantially lower than it had been in previous studies using this paradigm and was substantially higher than usual in the second (~44%). A base-rate of about 25% cheaters had been observed previously. So it is possible that knowledge of previous base rates may have actually made this task more difficult. Debriefing with the experts showed that this was the case. At least one of the experts was aware that their judgments departed from prior base-rates, and that this made them uncomfortable.

It is also important to note that our two experiments are part of a larger research program. The cheating paradigm has been utilized several times, numerous peer-reviewed papers have been published from the results, and our current results should be considered in this larger context. With Scripting questions, as observed in many prior published experiments, accuracy is not nearly so high (Levine et al., 2014). An accuracy of over 80% has never been obtained in any of our other studies using the cheating paradigm regardless of the questions asked so
long as the questions were asked by a researcher, not an expert. If Vrij et al. (2015) are right and the accuracy was because of the nature of the truths and lies, not who asked the questions and how, then this would not be so. It follows from Vrij et al.’s assertions that the cheating paradigm would typically and invariably yield high accuracy and high rates of confessions. It does not. Results have been published using truths and lies from the cheating paradigm with deception detection accuracies for experts as low as 17% (Levine et al., 2011). In short, other published studies using the cheating paradigm empirically demonstrate that Vrij et al.’s premise, that it is the nature of the truths and lies that produced the results, is false (Levine et al., 2014).

Many of the subjects, cheaters and non-cheaters alike, reported that the experiment was the most intense communication experience of their lives. It was nothing like the usual deception task with scripted questions. It was hard for most liars to maintain their lies, but this is not because they could not make up an explanation for a question or two. Lying was hard because of the intense probes that followed; where every inconsistency was explored, and where personal morality of the interviewee was directly challenged. It takes much training and experience to be able to communicate the way our experts did, and maintaining a credible lie through questioning that intensely is indeed difficult. As our results showed, the vast majority of guilty subjects could not; but, that was because of the nature of questioning, not the nature of the cheating set-up.

Vrij et al. (2015) also critique the confessions in our study. They argue that the answers to the “how did you know the answers” question was so diagnostic that the investigators never had to apply the coercive strategies that produce false confessions. This is false. These “coercive” strategies were utilized in every interview in Study 1. The interviewer used minimization/maximization techniques and false evidence in every interview during study 1. These are the same techniques that regularly produce false confessions in other paradigms. Confessions were highly diagnostic in this study only because the experts were skilled at avoiding false confessions and at soliciting honest confessions. Other iterations of the cheating experiment never yielded such a high rate of confessions, and the experts could not know the proportion of cheaters who would confess. The exceptionally high confession rate in the two expert experiments relative to non-expert interviews in the same cheating paradigm is strong evidence that it was what the experts did that mattered.

Vrij et al.’s misrepresentations of our method mystify us. Vrij et al. assert that the content of the questioning (i.e., the questions and the answers) was not disclosed. While it is true that there were no transcripts in the article, as they well know, videotapes were made of all our interviews from various cheating experiments and are available for academic use (see http://timothy-levine.squarespace.com/deception-interviews/). Furthermore, two of our critics (Vrij and Meissner) are part of a research consortium for which complete transcriptions were provided and Meissner was involved with our funding. Thus, our critics had access to our data, and they could have requested clarification or the actual video tapes. One of our critics (Meissner) presumably evaluated our design before the data were collected in connection with our funding but he does not disclose this. Further still, the article in question is part of a larger program of research. It is clear Vrij et al. (2015) know this because they cited works of ours that clearly contradict their own premises. In any case, the concerns they criticize are not present in the research program that is the topic of their criticism.

In summary, the Vrij et al. (2015) critique is based upon five false premises. These are: (1) honest cheaters knew and could convincingly explain how they knew the answers; (2) liars could not explain how they got the answers correct; (3) honest interviewees could prove their innocence; (4) false confessions were not obtained because no coercive tactics were used; and (5) a simple decision rule can explain the results. Premises 1 through 4 are simply and verifiably false with video evidence from our experiment. Simple probability theory was used to show that the fifth premise is implausible as an explanation. The fifth premise, also, requires hindsight to function. Our experts did not possess hindsight and thus any premise that requires hindsight as an assumption must also be false. Because Vrij et al. (2015) argument is based on a string of false premises, their conclusion cannot be sound. Sound argument and sound scientific reasoning requires premises that are true.

The diagnostic utility of confessions

What if we had a simple decision rule; people who confess are guilty; otherwise guilt is uncertain? Vrij et al. correctly note that false confessions can and do occur. They have led to wrongful convictions. Instances are well documented. This fact is not disputed. What percent of confessions to previous lies outside the lab are false? In the context of deception detection, is accuracy improved or hindered by trying to persuade a potential liar to be honest? We believe that the mere presence of some non-zero false positive rate does not justify the conclusion that the strategy is so dangerous that it must be banned.

Consider health screening as an analogy. Almost all screening tests have a non-zero false positive rate. When a false positive occurs, it can cause substantial emotional and physical distress to the patient. This distress can include the performance of unneeded and even potentially fatal procedures; yet, we continue to use these tests because, on balance, the positives outweigh the negatives. If we extend Vrij et al.’s argument to its logical conclusion, we would declare almost all health screening tests dangerous and ban them.
It may well be that the dangers of a false confessions outweigh the benefits of pursuing true confessions, but several key pieces of information are being missed to make that determination. Among these are: how likely are false confessions relative to honest confessions under field conditions, can the risk be mitigated through other procedures, how many guilty criminals would be allowed to go free if interrogations were not allowed, and what damage would they do to society? Vrij et al. may believe that they have the answers to some of these, but we believe their information is lacking, and without solid information, it is not possible to develop informed policy. Without this knowledge, the argument is simply that the possibility for error exists, and therefore, it should never be done.

Our data showed it is possible to persuade cheaters to be honest without producing false confessions. This does not mean that false confessions cannot happen. Obviously they can and do. The fact that they did not happen in our study, however, does not make our findings dangerous, it makes our findings promising.

When experts perform well (and not)

Deception detection experts almost always perform poorly in deception detection experiments. Let us compare the deception detection task in the typical experiment to the task faced by a skilled investigator. Real criminals have motives. Motives are irrelevant in most deception detection experiments where students are randomly assigned to conditions and are just following researcher instructions.

Real investigators usually have some evidence and much contextual knowledge. Evidence and contextual knowledge are irrelevant in most deception detection experiments. Communication content matters in most real interrogation situations. In the lab, cue-based detection is usually the only option. Content has little utility in controlled experiments. Outside the lab, an investigator can use persuasion to convince an interviewee to be honest. In the lab, subjects who confess are removed for a failure to follow instructions.

We believe that experts perform poorly in deception detection experiments, in part, because the experiments do not simulate the environment in which the expertise develops and limit available strategies to those that lack diagnostic utility.

In our experiment, liars were hiding some bad behavior. Projecting motive (Levine et al., 2010) became a viable strategy as did assessing an individual’s integrity. While the investigator did not have hard evidence, they were familiar with the context and communication content was relevant (Blair et al., 2010).

Interviewers were allowed to try and persuade cheaters to be honest. Under these conditions, experts did quite well when allowed to ask their own questions - as they would outside the lab.

Conclusion

We recently published two experiments yielding exceptionally high deception detection accuracy. Although we anticipated higher than usual accuracy, our results surprised even us. We made sure to replicate the findings in a second experiment before we sought to publish the results. Even then, we expected skepticism from the research community. If it weren’t our own data, we’d be skeptical too.

Vrij et al. (2015) wrongfully dismiss our findings. Our experts, they claim, were given a trivially easy task. It is incumbent on those who would explain our findings to base their explanations on what was done and not some caricature of our paradigm. The task was not so easy. The number of trivia questions a subject got correct in the game was highly diagnostic, but not so much as to explain our results. The plausibility of explanations were also diagnostic, but not deterministically so. The critics of our work have hindsight knowledge that our experts lacked. Our experts found the task quite challenging and so did the interviewees. Although the experts did exceptionally well, none thought doing so was simple or easy.

In the past, researchers have put experts in a lie-detection environment where they could not perform well and they did not. Experts in our two experiments were put in a context where high performance was possible. Our results are no more an artifact of our method than the results of any other study. The proper conclusion is that cue-based lie detection is poor while content-and-persuasion-based lie detection is promising.

This said, the experts in our experiment were elite. They had extensive training and experience. Our results do not show that all experts are good lie detectors only that some experts can perform well under certain circumstances. Some of the same experts have been used in cue-based lie-detection tasks (Blair et al., 2010; Levine et al., 2011; Levine et al., 2014) and they did not perform nearly as well when watching cheating tapes using scripted questions, even when the questions asked included the number correct and how the participants knew the answers. We think soliciting honest admissions is a viable approach, and it certainly worked well for the experts in our experiments. We look forward to seeing how our ideas and research will hold up as knowledge progresses.

CONFLICT OF INTERESTS

The authors have not declared any conflict of interests.

REFERENCES
