A critical look at meta-analytic evidence for the cognitive approach to lie detection: A re-examination of Vrij, Fisher, and Blank (2017)

Timothy R. Levine1*, J. Pete Blair2 and Christopher J. Carpenter3

1University of Alabama Birmingham, Alabama, USA
2Texas State University, San Marcos, Texas, USA
3Western Illinois University, Macomb, Illinois, USA

Purpose. This article provides a re-analysis of Vrij et al.’s (2017, Leg. Crim. Psychol. 22, 1) meta-analysis of the cognitive approach to lie detection. Vrij et al.’s analyses confounded dependent variables, capitalized on aberrant controls, and used unreliable data to inflate support.

Methods. Meta-analysis was used to reanalyse Vrij et al.’s data. Studies of human detection and studies involving statistical classification were analysed separately.

Results. The advantage offered by the cognitive approach was much smaller than previously claimed. Accuracies in control conditions were unusually low, and the most supportive findings came from the least reliable data.

Conclusions. Human detection and statistical classification are different. The evidence for the cognitive approach has been overstated.

Vrij, Fisher, and Blank (2017) published the first meta-analysis assessing the effectiveness of Vrij’s (2015) new cognitive approach to lie detection. Their basic premise is that verbal and nonverbal deception cues can be amplified by making lying more cognitively effortful. Cues can be amplified in three ways: (1) instilling additional cognitive load, (2) prompting additional information, and (3) employing unexpected questions. Amplified cues lead to more accurate deception detection.

Vrij et al. (2017) reported that their cognitive approach showed a 15-point advantage over the traditional standard approach. They claimed 71% correct classification for their cognitive approach compared to 56% for the standard approach in head-to-head experimental comparisons.

This essay provides a critical look at Vrij et al.’s (2017) claims, methods, and data. Vrij et al.’s meta-analysis averaged across two conceptually, empirically, and pragmatically different types of outcomes. As a result of conflated outcomes, evidence for the cognitive approach was exaggerated and important patterns of findings were hidden. Specifically, accuracy in the controls and reliability in assessment correlate negatively with support for the cognitive approach.

*Correspondence should be addressed to Timothy R. Levine, Department of Communication, University of Alabama at Birmingham, Birmingham, AL 48824-1212, USA (email: levinet111@gmail.com).

DOI:10.1111/lcrp.12115
Confounded outcomes

Meta-analysis is a family of statistical procedures used to combine data from multiple prior studies called primary studies. If the variables differ conceptually from study to study, then the primary studies necessarily involve different kinds of effects. Different kinds of effects cannot be meaningfully averaged into a single effect. When different effects are averaged as if they were the same thing, the findings are said to be confounded.

The Vrij et al. (2017) meta-analysis aggregated two different sorts of outcomes. Some of the primary studies looked at how human subjects perform in lie detection tasks. Other primary studies involved ‘objective criteria, using, for example, a discriminant analysis algorithm’ (Vrij et al., p. 8). In this second type of study, researchers coded behaviours (cues) of subjects who lied or told the truth and used a statistical algorithm to classify subjects’ honesty–deceit based on the researcher-coded behaviours.

Human judgement of the honesty is different from statistical classification based on researcher-coded behaviours and statistical analysis. In human detection studies, the subjects are human receivers who make decisions about senders. In the statistical classification studies, the subjects are senders and the outcomes come from computer programs. The computer findings tell us about behavioural cues that differentiate between truths and lies. The human judgement studies tell us about how well humans can differentiate between truths and lies – presumably based on behavioural cues. Human judges do not have the benefit of ground truth hindsight or lots of data for computing fit, much less the cognitive capacity and know-how to perform statistically optimized classification in their heads in real time.

In the logic of the cognitive approach, the treatment enhances behavioural cues which may subsequently enhance human judgement to the extent that humans use those enhanced cues. That is, to the extent to which the cognitive approach is valid, behaviours mediate the effects of the cognitive approach on human judgements.

Vrij et al. (2017) consider these two types of outcomes (human judgement and statistical cue classification) as a potential moderator. Outcome type did not moderate treatment effects on total accuracy. To quote Vrij et al.: ‘How the veracity decisions were made (by humans or based on objective criteria) did not matter very much for total accuracy, $Q_H(1) = 0.85, \ p = .36$’ (Vrij et al., 2017, p. 8). Vrij et al. then averaged across the two types of studies. The 71% and 56% averages they reported for total accuracy included both human and statistical classification results. These percentages are confounded.

Empirical data strongly refute Vrij et al.’s (2017) claim that outcome type does not matter for total accuracy. Although outcome did not moderate the treatment effects, statistical classification consistently produces much higher accuracy than human judges. This has to be the case because we know that human judgement is imperfect (cf. lens model analyses of deception detection, Hartwig & Bond, 2011). Computer outcomes tell us about the link between honesty–deception and behaviours. Human judgements involve two links in a causal chain: deception to behaviours and behaviours to judgement. To the extent that human judgement deviates from perfection, human outcomes must be lower than outcomes based on objective behaviours (unless the findings are inflated by methodological artefacts as we later argue).

In Table 1, weighted means were calculated based on our re-analysis of Vrij et al.’s (2017) data and compared to prior meta-analyses involving human judgement (Bond & DePaulo, 2006) and cue-based statistical classification (Hartwig & Bond, 2014). Human observer truth–lie discrimination is 14–15 points lower than objective criteria using the
standard approach in the Vrij et al. meta-analysis, 19 points lower in prior meta-analyses, and 16–20 points lower in the Vrij et al.-analysed cognitive approach conditions.

Because human lie detection tasks and statistical classification algorithms are so theoretically, practically, and empirically different, they are not apples-to-apples units amenable to meaningful summation in meta-analysis. Instead, they are apples and oranges. Human lie detection and objective criteria classification need to be understood as qualitatively different outcomes from different sets of studies. Separate analyses are needed.

### Capitalizing on aberrant controls

The confounded outcomes masked the existence of aberrant controls that undercut the internal validity of the experimental comparisons. The following quote from Vrij et al. (2017) demonstrates how the aggregation of human and statistical classification is used to draw misleading conclusions:

> It is noteworthy that the 56% (total) accuracy rate obtained with a standard approach very closely approximates the typical 54% accuracy rate typically found in deception research (Bond & DePaulo, 2006), $z = 1.08, p = .28$. By contrast, the 71% total accuracy rate obtained using the cognitive lie detection approach significantly exceeds this 54% accuracy benchmark, $z = 9.37, p < .001$(p. 6)

The quote above makes the dual claims that accuracy in the controls (standard approach) was typical (56% vs. 54%) while accuracy with cognitive lie detection is significantly higher than the average obtained in a prior meta-analysis (71% vs. 54%). However, the Bond and DePaulo (2006) meta-analysis reporting 54% accuracy included only human detection findings. The 56% and the 71% to which Vrij et al. point contained both human and statistical classification. It follows that the 56% and the 71% in the Vrij et al. quote are not comparable to the 54% in Bond–DePaulo because different outcomes are being compared. The apples-to-apples comparisons are, for human observers, the 48–49% accuracy for human detection in standard (control) approach in Vrij, the 54% in Bond–DePaulo, and the 58–62% for the cognitive approach. For objective criteria, the comparison is 63% for the standard (control) approach in Vrij et al., 73% for the standard approach in Hartwig and Bond (2014), and 78% for the cognitive approach in Vrij et al.
The gain relative to prior meta-analysis is 4–8 points rather than the 17 points claimed (see Figures 1 and 2).

We call this idea Capitalization on Aberrant Controls. If any two means are statistically different, the difference can be because the value of the lower mean is depressed, the value of the larger mean is elevated, or both. The benefits of the new treatment are illusory, and claims of improvement are specious when a statistical difference is a function of unusually poor performance in the control group. An example of unusually poor performance in controls was pointed out with regard to research on unconscious lie detection (Levine & Bond, 2014).

To clearly see the issue, we need to be explicit about what is expected in control groups. Both human judges and objective criteria research have been extensively studied for decades and have been summarized in competent and comprehensive meta-analyses. For human detection, Bond and DePaulo (2006) is the preferred meta-analysis. As Vrij et al. (2017) note, the across-study average accuracy for human lie detectors is 54%. But, understanding what is aberrant also requires knowing how findings from individual
studies are usually distributed. Bond and DePaulo show that standard findings are normally and tightly distributed around 54%. Almost 60% of the prior findings were in the 50–60% range. Less than 14% of prior findings were below 48% accuracy and only 5% fell below 45%. Accuracy in the 40s happens, but it is unusual relative to findings in the 50–60% range at a ratio of 1:4. Further, because the distribution approximates normality, the distribution of expected findings is symmetrical. Accuracy at or below 48% occurs at about the same rate as accuracy at or above 60%. Thus, if controls in the Vrij et al. meta-analysis reflected the standard approach, we would expect a majority of findings involving human judges to fall in the 50s and the minority of findings outside the normative range to be equally split between findings in 40s and 60s.

The findings of the primary studies with human judges included in Vrij et al. (2017) are summarized in Table 2. In the seven experiments for which a standard control is reported, four reported values below 48% and two reported accuracy below 45%. In the controls, accuracy in the 40s or below is more common than accuracy in the normative 50–60% range, and not a single primary study reported accuracy at 60% or above in the controls. These findings do not comport with the larger literature.

Next, we consider the normative performance of statistical classification with objective criteria. Hartwig and Bond (2014) provide the relevant meta-analytic comparison. Average statistical classification accuracy based on objective measures in the literature is 73%.1 Examination of the findings listed in Vrij et al.’s appendix reveals that every primary experiment included in their meta-analysis had a result in the control group that was below the 73% average typical of the larger literature.

In all, five of seven human observer comparisons and seven of seven objective measure comparisons (12 of 14 overall) show performance in the control conditions

<table>
<thead>
<tr>
<th>Study</th>
<th>Standard control</th>
<th>Cognitive approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evans, Meissner, Michael, and Brandon (2013, Study 1)</td>
<td>37&lt;sup&gt;a&lt;/sup&gt;</td>
<td>67&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Vrij et al. (2008)</td>
<td>46&lt;sup&gt;a&lt;/sup&gt;</td>
<td>58&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>Vrij, Mann, Leal, and Fisher (2010)</td>
<td>52&lt;sup&gt;c&lt;/sup&gt;</td>
<td>54&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>Vernham et al. (2014)</td>
<td>47&lt;sup&gt;a&lt;/sup&gt;</td>
<td>77&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Zimmerman et al. (2010)</td>
<td>42&lt;sup&gt;a&lt;/sup&gt;</td>
<td>58&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>Colwell et al. (2009)</td>
<td>—</td>
<td>67&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Liu et al. (2010)</td>
<td>57&lt;sup&gt;c&lt;/sup&gt;</td>
<td>59&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>Vrij et al. (2011)</td>
<td>56&lt;sup&gt;c&lt;/sup&gt;</td>
<td>69&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

<sup>Note</sup>. <sup>a</sup>Example of an aberrant control.  
<sup>b</sup>Apparent evidence for substantially improved accuracy.  
<sup>c</sup>Value within the 50–60% typical of the literature.

---

1 The Hartwig and Bond (2014) classification accuracy is not downward-corrected for cross-validation and was obtained from Charlie Bond on 10/14/15. The 68% classification accuracy reported in the Hartwig and Bond article is an estimate of cross-validated classification accuracy. We use a non-corrected estimate to avoid confounding as most findings included in Vrij et al. (2017) were not cross-validated.
that is below average in comparison with the larger literature as summarized by meta-
analysis. At least half of the advantage claimed by the cognitive approach is
attributable to systemic deflation in the controls.

Vrij et al. (2017) provide a clear example of the capitalization of aberrant
controls. They consider only the improvement provided by the new cognitive
approach, even though their data show that much of the difference they tout is
due to unexplained and unusually poor performance in the standard control
conditions. This is true both for the human judgement findings and the objective
criteria studies. However, Vrij et al. not only fail to mention this finding; they
explicitly deny it (see the quote above). The way the data are reported functionally
disguises the aberrant controls by confounding human and objective criteria
outcomes.

The prevalence of aberrant controls raises concerns about the processes that create
the aberrant controls. Two possible candidates are researcher degrees of freedom and
publication bias. Unless these can be ruled out, research consumers have good reason to
be sceptical of the findings summarized in Vrij et al. (2017).

A re-analysis of Vrij et al. (2017)

We collected and read each of the primary studies included in the Vrij et al. (2017)
meta-analysis. We confined our re-analysis to the primary studies involving head-to-
head comparisons of some standard to the cognitive approach. The statistical data
needed for meta-analysis were independently extracted from the primary articles by
two of the current authors. Three-way comparisons were made between values coded
by the current authors and the Vrij et al. report to ensure the accuracy of the data.
Discrepancies were factually resolved by double-checking primary articles. All
discrepancies were attributable either to rounding or ambiguity in selection among
multiple findings. Rounding errors were corrected. When multiple findings were
available for inclusion, the values chosen by Vrij et al. (2017) were also used in our re-
analysis. Because our goal was an exact replication, we abided by Vrij et al.’s
inclusion decisions.

Our examination of the primary studies raised additional concerns regarding the
meaningfulness of the Vrij et al. (2017) meta-analysis. First, it is debatable if the
collection of findings provides conceptually coherent samples of some standard
approach that can be meaningfully compared with samples of a coherent alternative
cognitive approach. Examination of what was counted as the standard approach
revealed a diverse assortment of tasks including accuracy stemming from opening
questions in an interview protocol (Vrij et al., 2009), a structured interview protocol
‘designed to maximize recall and minimize contamination’ (Colwell, Hiscock, &
Memon, 2002, p. 289), a constrained variety of questioning techniques and interview
approaches currently used by the United States Army (Zimmerman, Veinott, Meissner,
Fallon, & Mueller, 2010), having a second silent interviewer with a neutral or
suspicious demeanour (Mann et al., 2013), and asking questions about past activities
(rather than intentions, Vrij, Leal, Mann, & Granhag, 2011).

In contrast, the cognitive approach included asking event-irrelevant questions
(Liu et al., 2010), asking spatial, temporal, and drawing questions after opening
questions (Vrij et al., 2009), providing a model statement for senders (Leal, Vrij,
Warmelink, Vernham, & Fisher, 2015), having a second silent interviewer with a
supportive demeanour (Mann et al., 2013), forced changes of two interviewees who were speaking in 20 s intervals combined with instructions to judges to look for specific cues identified by coders (Vernham, Vrij, Mann, Leal, & Hillman, 2014), and asking questions about intentions (rather than past activities, Vrij et al., 2011). To our eyes, both sets look to be diverse collections of methods and findings rather than two internally coherent and conceptually different approaches. We could not independently reproduce Vrij et al.’s (2017) coding regarding what was standard and what fit under the cognitive approach. As with our decisions regarding the selection of which findings to include, our re-analysis gave Vrij et al. the benefit of doubt. That said, it is our opinion that the collection of primary studies included in Vrij et al. are unsuitable for meta-analysis because they lack discernable conceptual coherence.

Our second major reservation stems from the quality of research exemplified in the primary studies. To our eye, the most important issue in the primary studies was the unreliable assessment of the outcome measures in the human detection studies. Single-item assessment was evident in over 40% of the human detection experiments. This is absolutely critical because prior meta-analysis (Bond & DePaulo, 2006) showed that human accuracy findings stabilize somewhere between 1,500 and 3,000 judgements (where judgements equals the number of subjects times the judgements per subject). In the human observer primary studies included in Vrij et al., the median number of judgements per condition is 67. As a consequence, the primary data involved very noisy estimation. Precision is paramount in scientific observation. Further, as shown in our re-analysis, what should be random noise shows a pattern of bias towards higher accuracy in the treatment conditions.

A related concern exists with the primary studies involving statistical classification based on objective criteria. Classification accuracy is based on some weighting of the objective behaviour(s) that statistically optimize the result for the known outcome (truth or lie). In interpreting statistical classification accuracy, it is important to understand that all data sets will have random error and other idiosyncrasies. This is especially the case with small sample data. As a consequence, classification accuracy statistically optimized for some particular data will be inflated and not projectable beyond the specific data upon which the classification was performed. To the extent that a classification model is based on the idiosyncrasies in some particular data instead of the ‘true’ relationship among variables, a model is said to be ‘overfit.’ To assess the utility of a discriminant function beyond the unique data from which it was derived, cross-validation is required. Unfortunately, only one primary study (Köhnken, Schimossek, Ascherman, & Höfer, 1995) involved any type of cross-validation. In no case was a classification model tested on new data.

Notwithstanding these and other concerns, we nevertheless conducted a re-meta-analysis of the Vrij et al. (2017) data using the Schmidt and Hunter (2015) approach, $r$ as the effect size, and weighting human judgement results by both the number of subjects and the number of observations ($n \times$ judgements per subject). We weighted means both ways to see whether our concerns regarding noisy estimation made a difference in the results beyond mere instability.

Across human and objective measures outcomes, the difference between the accuracies in the standard approach and the cognitive approach conditions produced a weighted mean effect of $r = .185$, with a 95% confidence interval of 0.115–0.254 and an 80% credibility interval of 0.079–0.291. Converted, the effect size was $d = 0.377$. 
Credibility intervals indicate how much the effect size is likely to vary due to undetected moderators. Because zero is outside the confidence interval, the treatment effect is statistically significant. However, only 64% of the study-to-study variation was attributable to sampling error suggesting heterogeneity of effects and likely moderators (Schmidt & Hunter, 2015).

Vrij et al. (2017) reported Cohen’s $d = 0.42$ with 95% confidence interval of $0.26 < d < 0.58$ and significant heterogeneity. Using the standard conversion, the $d = 0.42$ reported by Vrij et al. corresponds to $r = .205$ (CIs 0.13–0.28). While our analysis did not produce numerically identical results, the overall results are quite similar and yield the same substantive conclusion. Taken at face value, the prior research included in Vrij et al. show that the cognitive approach produces a statistically significant but modest improvement over the standard approach with effect sizes in the small to moderate range.

Although both Vrij et al. (2017) and our re-analysis produce nearly identical results, a point of interest is the metric in which the findings are presented. Vrij et al. focus on raw within-cell accuracy rather than the effect sizes. They summarize their findings like this:

> In sum, the meta-analytic findings confirmed the superiority of the cognitive lie detection approach. Specifically, the cognitive lie detection approach led to better total detection accuracy (71% vs. 56%) and also into more accurate classifications of truth tellers (67% vs. 57%) and liars (67% vs. 47%), when compared directly to a standard approach. (p. 10)

The results seem less impressive if the focus is on the effect size. We noticed that the effect size for the difference between the standard and cognitive approach in the Vrij et al. (2017) meta-analysis ($d = 0.42$) is identical to the difference in prior meta-analysis between 50 and 50 chance and the 54% across-study average ($d = 0.42$; Bond & DePaulo, 2006).

Next, we did separate meta-analyses for human and objective measures experiments. For human judges, the weighted mean effect was $r = .213$, with a 95% confidence interval of 0.076–0.350 and an 80% credibility interval of 0.020–0.406. Converted, $d = 0.436$. Only 33% of the study-to-study variation was attributable to sampling error suggesting massive heterogeneity. This finding provides clear statistical support for our contention that the human experiments fail to replicate (see Table 2).

For objective measures, in contrast, 100% of the variance in effect sizes was explainable in terms of sampling error suggesting a homogeneous set of findings that were sampled from a single population. The weighted mean effect was $r = .150$, with a 95% confidence interval of 0.118–0.199. Converted, $d = 0.303$.

We noticed that the logic of the cognitive approach predicts larger effects on behaviours than judgements. The approach seeks to enhance behavioural cues directly and judgements indirectly through behaviours. Yet, effects were larger (although not

---

2 Readers may wonder how the standardized difference associated with a 15-point swing in Vrij et al. (2017) is quantitatively the same as a 4-point swing in the Bond and DePaulo (2006) meta-analysis. The difference is the relative noisiness of the data. Effect sizes are ratios of effects to effects plus error. In Bond and DePaulo, there is a small difference (50% vs. 54%) divided by the small difference plus a relatively small amount of error. In Vrij et al., there is a much larger difference (56% vs. 71%) divided by this much larger difference plus much larger error (largely a function of inadequate measurement). Focusing primarily on the numerator allows Vrij et al. to simultaneously present their results in a way that makes their findings look big while hiding the noisy nature of data. Focusing on effect sizes ($r$ or $d$; not odds ratios) rather than significance tests and raw accuracies suggests that improvement provided by the cognitive approach is actually modest.
Finally, we examined weighted mean classification accuracy for the standard and cognitive approach broken down by human and objective outcomes. For the human outcomes, we weighted the results by both sample size and the number of judgements. For human outcomes, the standard approach yielded accuracy of 48.13% (SD = 40.00%, 95% CIs = 43.44–52.81%) when weighted by sample size and 48.99% (SD = 25.29%, 95% CIs = 47.73–50.26%) when weighted by number of judgements. One-sample t-tests show that both values are significantly below the prior meta-analysis mean of 54% documenting the aberrant nature of the findings in the controls; \( t(278) = -2.46, p < .05 \) and \( t(1,527) = -7.75, p < .05 \). These tests show that humans are significantly worse lie detectors when they are randomly assigned to control conditions in the literature examined by Vrij et al.

Regardless of how the findings are weighted.

The weighted averages for the cognitive approach conditions with human judges were 62.08% (SD = 57.56%, 95% CIs = 57.72–66.43%) when weighted by sample size and 57.78% (SD = 23.60%, 95% CIs = 56.61–58.94%) when weighted by number of judgements. One-sample t-test shows that both values are significantly above the prior meta-analysis mean of 54% documenting improvement relative to the literature on human lie detection as whole; \( t(284) = 3.64, p < .05 \) and \( t(1,573) = 6.36, p < .05 \). Both values, however, are significantly below the 71% claimed by Vrij et al. (2017); \( t(284) = 4.57, p < .05 \) and \( t(1,573) = 22.23, p < .05 \). Vrij et al. (2017)’s claim of 71% accuracy does not reflect human performance in their analysis.

Curiously, average human accuracy in the cognitive approach conditions depends on how the mean is weighted. The unweighted mean is 63.14%. That drops slightly to 62.08% when we weight by sample size. The average drops further to 57.78% when we weight by the number of judgements. This is because the strongest result for the cognitive approach (77%) was obtained in the least reliable study (based on a mere 30 judgements) while the weakest finding (54%) was based on the most (832) judgements. Looking at Table 2, we see that findings in the treatment conditions cluster into two distinct sets. On one hand, there exists a set of four experiments (67%, 67%, 69%, and 77%; average 70%, total \( N = 138 \) who made a total of 314 truth–lie judgements) that show substantial improvement for the cognitive approach. On other hand, for every strongly supportive finding, there is a contrasting finding (54%, 58%, 59%, 59%; average 57%, total \( N = 178 \) who made a total of 1,245 truth–lie judgements) showing only meagre improvement. The correlation between the reported accuracy for the cognitive approach is significantly so on judgements \( (d = 0.44) \) than behaviours \( (d = 0.30) \). This finding is inconsistent with the theory.

\[ \text{Our argument presumes that the untrained college students serving as human subjects do not somehow know more about valid lie detection using the cognitive approach than the researchers. If untrained college student judges based their assessments on cues with greater utility than the cues that the researchers chose to code, then human detection might yield stronger results than objective classification. Meta-analysis (Bond & DePaulo, 2006; Hartwig & Bond, 2014), however, finds much smaller effect sizes for human truth–lie discrimination (effect size \( d = 0.42 \), raw accuracy = 54%) than meta-analysis of objective behaviours (effect size \( R = 0.53 \), raw accuracy = 73%) even with single cues, \( r = .4, d = 0.87 \). Objective measures substantially outperform human detection just as theory dictates (see Table 1). Oddly, in Vrij et al.’s (2017) findings, this difference reverses. Given the nature of the primary studies included in Vrij et al. (2017), it is hard to understand how human judges could possibly match objective classification in effect size for an approach meant specifically to increase the utility of cues. Not only were human judges usually untrained, often times they made single assessments based on a single sender in a novel situation. Human judges were presumably blind to the independent variables and the specific purpose of the study. Objective measures studies, in contrast, had trained coders quantify between 2 and 19 behaviours (mean 8.7) selecting 1–16 of those behaviours (mean 5.9) for computing statistically optimal classification. Researchers coded the cues that should be most impacted by the experimental manipulations and picked from among multiple cues when creating the classification algorithms. Valid experimental tests of the cognitive approach would produce substantially larger effects on cues than human judgements.} \]
approach and number of judgements used to calculate accuracy in each primary study was $r = -0.59$, $p = 0.16$. That is, the more reliable the outcome, the lower the performance for the cognitive approach. The negative association between number of judgements and accuracy was not observed in the control conditions, $r = 0.10$. Oddly, what should be random noise stemming from unreliable single-item measurement tends to favour the cognitive approach.

While the accuracy observed in the cognitive approach conditions varied by the number of judgements, the treatment effects (in effect size $d$) in the human detection were a function of the accuracy in the controls. The correlation between the accuracy reported in the controls and the treatment effects ($d$) was $r = -0.81$, $p < 0.03$. We sorted the primary studies into two groups based on the accuracy reported in the controls. Three experiments reported controls that closely approximated typical results (52%, 56%, 57%, all within ±0.5 SDs of the prior meta-analysis mean of 54%). The average treatment effect for the cognitive approach (effect size $d$) in the three experiments with typical controls was $d = 0.14$, dropping to $d = 0.116$ when weighted by number of judgements. The other four experiments had aberrant controls in which accuracy was more than 1 SD below the expected mean based on prior meta-analysis (47%, 46%, 42%, and 37%). In the four experiments exhibiting usually low controls, the treatment effect was $d = 0.76$ (increasing to $d = 0.88$ when weighting by number of judgements). Clearly, the findings capitalize on aberrant controls and unreliable assessment.

For the primary studies involving objective measures, the weighted average for the standard approach was 62.96% ($SD = 48.63$, 95% CIs = 57.12–68.79%). This value was significantly lower than the 73% obtained in prior meta-analysis, $t(265) = 3.37, p < 0.05$. The weighted average accuracy for the cognitive approach using objective measures was 77.92% ($SD = 41.38$, 95% CIs = 73.14–82.70%), a value significantly larger than the 73% benchmark, $t(286) = 2.02, p < 0.05$. Similar to the pattern in the human observers’ data, the correlation between classification accuracy in cognitive approach and sample size was $r = -0.44$.

We tested our contention that human and statistical outcomes are statistically different. In the standard (control) conditions, human observers produced a sample-size weighted mean accuracy of 48.13% ($SD = 40.00$, $N = 280$) compared to 62.96 ($SD = 48.63$, $N = 267$) for objective classification, $t(544) = 3.90, p < 0.05$. For the findings in the cognitive approach conditions, human observers produced a sample-size weighted mean accuracy of 62.08% ($SD = 37.56$, $N = 286$) compared to 77.92 ($SD = 41.38$, $N = 288$), $t(566) = 4.80, p < 0.05$. Contrary to the claims of Vrij et al. (2017), their own data show that these are different outcomes.

**Summary of findings**

Figures 1 and 2 visually summarize our key points. Vrij et al. (2017) conflated studies of human judgement with results involving statistical classification of objectively coded behaviours (cues). Figure 1 presents the empirical story told by the human observer data, and Figure 2 provides the results from studies involving cue-based objective statistical classification. Comparing Figures 1 and 2, it is immediately apparent that the baseline expectations from prior meta-analysis are quite different for the two types of outcomes: 54% in human data and 73% the objective classification literature. Figures 1 and 2 both show that standard (control) findings fall well below the levels expected from prior meta-analysis and that the differences between the standard and the cognitive approach are
Conclusions

Vrij et al. (2017) claim that their cognitive approach produces an impressive accuracy of 71%, much superior to the 56% in the controls and the 54% in prior meta-analysis. These claims are confounded and inflated. The efficacy of the cognitive approach is more modest and magnified by aberrant findings in which processes that should be random do not appear random.

One issue was the confounding of human detection scores and statistical classification outcomes in the computation of total accuracy scores. Outcome type was included in Vrij et al.’s (2017) analysis, and it did not moderate the treatment effects on total accuracy. Total accuracy and treatment effects, however, are not the same thing. Vrij et al. emphasize total within-cell accuracy (the 71% claim), not effect sizes, in most of their article. The total within-cell accuracy results are indeed confounded. Our point is that although outcome type did not moderate the treatment effect (i.e., effect size), it does matter for raw accuracy scores (cell means) which is the primary way Vrij et al. presented their results to the readers. The conflation of accuracy types function to (1) make the results look much stronger than they really are and (2) hide some potential problems from view.

Theoretically, the two outcome types play different roles in the cognitive approach. The cognitive approach seeks to enhance behavioural cues. Human judgements are impacted indirectly as a consequence of the enhanced behaviours. Vrij et al.’s (2017) statistical treatment of outcome type as a moderator was not consistent with their theoretical argument in which behaviours are directly impacted outcomes and judgements are indirect outcomes. Further, their findings do not appear to be theory-consistent with respect to outcome types. If the cognitive approach was working as theoretically specified, treatment effects would have impacted behaviours more than judgements.

Our second issue was that accuracy was lower than expected in the controls. Somewhere between one-third and two-thirds (depending on the analysis and type of outcome) of the treatment effects (effect size) appears to be due to poor performance in the controls relative to expectations based on prior meta-analysis. We note that the aberrant controls were commonplace in both types of experiments (human and statistical classification). Why this is the case is unclear, but the impact is that the claimed effect sizes may be substantially inflated. Further, selection effects undercut the internal validity of the primary experimental comparisons.

The final major point of our critique involved that the dual observations that the primary studies contained extremely unreliable assessment due to relatively few judgements and that reliability was negatively associated with outcomes; that is, the most supportive findings were the least reliable, and the most reliable assessment produced the weakest findings. Why this is the case is again unclear. But the impact is that the claimed effect sizes are probably substantially inflated. Usually, of course, reliable data provide more precise estimates.

Putting this all together, our re-analysis of Vrij et al. (2017) showed a statistically significant improvement in accuracy for the findings included under the cognitive approach over the findings that served as controls. Taken at face value, the effect size partially a function of unusually poor performance in the controls indicating selection artefacts.
associated with the improvement is in the small to moderate range (approximately $d = 0.4$ or $r = .2$), about the same magnitude as the difference between 50% and 54% in prior meta-analysis (Bond & DePaulo, 2006). At least half of that effect, however, was attributable to inexplicably poor performance in controls. The effect is further qualified by a failure of human observer findings to replicate, the reliance on unreliable data, a lack of cross-validation in objective behaviour studies, and debatable coding decisions regarding what to count as treatment or controls. Given all this, scepticism is surely warranted. Claims for a 15-point improvement by the cognitive approach are not credible.

References


*Received 16 November 2016; revised version received 1 August 2017*