Commentary

Psychological Science
2015, Vol. 26(10) 1646–1648
© The Author(s) 2015
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797615597333
pss.sagepub.com

No Evidence for Unconscious Lie Detection: A Significant Difference Does Not Imply Accurate Classification

Volker H. Franz[1] and Ulrike von Luxburg[2]
[1]Department of Psychology and [2]Department of Computer Science, University of Hamburg

Received 4/22/15; Revision accepted 7/2/15

In 2014, ten Brinke, Stimson, and Carney reported that unconscious processes detect liars better than conscious processes, and that the success of conscious processes was typically close to chance (~54% correct; Bond & DePaulo, 2006). They concluded that "although humans cannot consciously discriminate liars from truth tellers, they do have a sense, on some less-conscious level, of when someone is lying" (p. 1103) and argued that "accurate unconscious assessments are made inaccurate either by consolidation with or correction by conscious biases and incorrect decision rules" (p. 1104). In short, ten Brinke et al. suggested that humans unconsciously know quite well whether someone is lying; however, conscious deliberations render these accurate unconscious assessments inaccurate.

Such conclusions could potentially have far-reaching practical consequences. For example, on the basis of these conclusions, one could advise jurors and eyewitnesses in court to rely mainly on their intuition and to avoid conscious deliberations. However, this is a dangerous road to travel. There are well-documented cases in which eyewitnesses erred in their intuitive judgment, and only conscious deliberation led to the truth (Loftus, 2003). Therefore, before concluding that "accurate lie detection is, indeed, a capacity of the human mind, potentially directing survival- and reproduction-enhancing behavior from below introspective access" (ten Brinke et al., p. 1104), we should make sure that there is strong scientific evidence. Although the plausibility of these data has already been challenged (Levine & Bond, 2014; but see ten Brinke & Carney, 2014), we show that the statistical reasoning ten Brinke et al. used is flawed and that a more appropriate analysis of their data does not provide evidence for accurate unconscious lie detection.[1]

Reanalysis of Data from Experiment 2

In Experiment 2[2] of ten Brinke et al., participants watched 12 videos of interrogations, 6 showing a liar and 6 showing a truth teller (participants were not told which was the liar or truth teller). Then participants performed two tasks. In the *direct* (*conscious*) *task*, they saw pictures of the suspects and classified them as liars or truth tellers; performance was essentially at chance level (49.6% correct; chance level is 50%). In the *indirect* (*unconscious*) *task*, the pictures of the suspects (*primes*) were masked, such that they could not be perceived consciously. Participants sorted visible words (*targets*) such as "deceitful" or "honest" into the categories of "lie" or "truth." Participants were significantly faster if prime and target were congruent (e.g., the word "deceitful" was preceded by a picture of a liar) than if they were incongruent. On the basis of this significant congruency effect, ten Brinke et al. concluded that there are "accurate unconscious assessments" (p. 1104) of liars versus truth tellers in the indirect (unconscious) task and that these assessments are better than the chance-level performance in the direct (conscious) task.

However, this conclusion is flawed. The test for a significant congruency effect is concerned only with the question of whether a true difference in response times (RTs) exists in the population, regardless of its size. One can conclude from this effect only that some classification of the suspects has happened, but the accuracy of this classification and whether it was more accurate than in the direct task remain unknown. To make the claim that the RTs are evidence for good unconscious classification, ten Brinke et al. would have needed to show that the RTs can be used to classify whether the suspects were truth tellers or liars. Only then would it be possible to compare the accuracy of this indirect classification with that in the direct task.

How can such an indirect classification be performed? Because of the experimental design, classifying suspects

Corresponding Author:
Volker H. Franz, General Psychology, University of Hamburg, von Melle Park 5, Hamburg 20146, Germany
E-mail: volker.franz@uni-hamburg.de

as truth tellers or liars is equivalent to classifying trials as congruent or incongruent (e.g., if a trial is classified as congruent and the visible target was deceitful, then the suspect is classified as liar). Because ten Brinke et al. argued that the congruency effect is evidence for accurate unconscious classification (fast RTs in congruent trials, slow RTs in incongruent trials), all one needs to do is find an appropriate threshold *t* and classify all trials with RTs smaller than *t* as congruent and trials with RTs larger than *t* as incongruent.

We performed this classification on ten Brinke et al.'s data using three different methods of calculating a threshold. First, under the assumption that RTs follow normal or log-normal distributions (Ulrich & Miller, 1993), the threshold that leads to the best expected accuracy in a design with an equal number of congruent and incongruent trials is the median of the RTs (e.g., MacKay, 2003; p. 190). Therefore, we classified the trials using within-participant median RT, computed the accuracy over all trials of the participant, and averaged the accuracies across participants. This resulted in an average accuracy of 50.6% (*SD* = 2.65).

Second, to avoid assumptions about the RT distributions, we selected a threshold according to the standard procedures of machine learning (Shalev-Shwartz & Ben-David, 2014): We randomly split the trials of each participant into equal-sized training and test sets (we reached similar results when we used other split sizes). On the training set, we determined the threshold that led to the best accuracy and used it to classify the test set. We repeated this procedure 10 times with different random splits of the data for each participant. This led to an average accuracy of 49.5% (*SD* = 2.60).

Third, to construct an overly optimistic upper bound (i.e., the highest accuracy that possibly could be achieved for the given data), we evaluated the accuracy of all possible thresholds over all trials of each participant, determined the best result, and averaged the obtained accuracies across participants. This resulted in an accuracy level of 53.7% (*SD* = 1.99), which means that for these data, no possible classifier exists with an accuracy larger than 54%—the very number that was interpreted as "detection incompetence" by ten Brinke et al. (p. 1098). In short, the classification accuracy in the indirect task is just as poor as in the direct task and for all practical purposes can be considered as being at chance level. There is no evidence for accurate unconscious assessments.

Although this result seems clear and consistent, one might ask whether it is fair to assess indirect classification performance with RTs from single trials. One might argue that it would be better to average the RTs of multiple trials, thereby reducing measurement error and possibly improving accuracy. It is known from machine learning that such procedures can under certain conditions improve accuracy (e.g., Bühlmann, 2004). We tested this idea following two methods: First, for each participant, we averaged all trials related to each suspect, classified these averages using the median across all suspects as the threshold, and averaged the accuracies over all participants. This resulted in an accuracy level of 51.9% (*SD* = 16.30). Second, we averaged the RTs related to each suspect across all trials and all participants, thereby including all available information for the classification. This resulted in an accuracy level of 50.0%. In short, even if we combined RTs from multiple trials and multiple participants, there is no evidence for better accuracy in the indirect task than in the direct task.

To understand why a significant difference does not indicate accurate classification, consider an intuitive example. Suppose one tried to classify individual adults as female versus male on the basis of their weights. The weight distributions for the two genders overlap a lot, so performance would be poor.[3] On the other hand, if one performed a standard *significance test*, one would form two groups of same-gender adults and compare their mean weights. If the groups were large enough, one could easily obtain a significant difference. This shows how a significant difference can coexist with essentially chance-level classification accuracy and that good classification accuracy of the individual adults cannot be inferred from the fact that the group means are significantly different. Good classification requires more: a clear separation of the female and male weight distributions at the level of the individual adults (or, in the ten Brinke et al. study, the RT distributions at the level of the individual suspects). We have shown that—no matter how we aggregated the individual trials from ten Brinke et al.—the RT distributions of liars and truth tellers always overlapped so heavily that no good classification could be obtained.

Conclusion

The data of ten Brinke et al. do not provide any evidence for accurate unconscious lie detection. A significant difference in the indirect task does not indicate accurate unconscious classification. For a meaningful comparison of the classification accuracies in the indirect and direct tasks, both tasks have to measure just that: classification accuracy. We thank ten Brinke et al. for making their data available (Eich, 2014). Such public access to data allows for a rapid self-correction of science without going through lengthy replication attempts first—which, even if successful, can easily take years (Ioannidis, 2012).

Editor's Note: The journal's Editor in Chief (Eric Eich) invited ten Brinke et al. to respond to this Commentary by submitting a rebuttal that would be subject to review. In reply, ten Brinke et al. wrote: "We appreciate the opportunity to respond to this Commentary but will respectfully decline. We hope that methodological and mathematical experts will take up this conversation to refine the methods available for comparing conscious and nonconscious classification in future research."

Author Contributions

V. H. Franz discovered the methodological flaws in ten Brinke, Stimson, and Carney (2014) and reanalyzed the data in the R software environment. U. von Luxburg verified all arguments and reimplemented the analyses independently in MATLAB. Both authors wrote the manuscript.

Acknowledgments

We thank Gilles Blanchard and Frank Rösler for comments.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Funding

U. von Luxburg was supported by Deutsche Forschungsgemeinschaft Grant LU1718/1-1, Research Unit 1735, "Structural Inference in Statistics: Adaptation and Efficiency."

Open Practices

All materials have been made publicly available via Open Science Framework and can be accessed at https://osf.io/7825t. The complete Open Practices Disclosure for this article can be found at http://pss.sagepub.com/content/by/supplemental-data. This article has received the badge for Open Materials. More information about the Open Practices badges can be found at https://osf.io/tvyxz/wiki/1.%20View%20the%20Badges/ and http://pss.sagepub.com/content/25/1/3.full.

Notes

[1] We implemented all analyses twice, once in the R software environment (Version 3.2.1; R Development Core Team, 2015) and once in MATLAB (The MathWorks, Natick, MA); the code used in R is available at https://osf.io/7825t.

[2] We concentrated on Experiment 2 because only in this experiment were unconscious stimuli presented. Experiment 1 investigated whether consciously visible pictures of the suspects had indirect effects on another task. Nevertheless, our critique also applies to Experiment 1, because ten Brinke et al. also inferred good classification from a significant RT difference for these indirect effects. However, our analysis—using the first threshold-calculation method that we discuss later—resulted in classification accuracy of 51.1% (*SD* = 3.71), which is again clearly below the 54% that ten Brinke et al. described as "detection incompetence" (p. 1098).

[3] We thank an anonymous reviewer for suggesting this example. The average weight difference between men and women is about 13 kg (*SD* ≈ 33 kg; McDowell, Fryar, Ogden, & Flegal, 2008), so classification accuracy would be about 58%, $d = 0.4$ (13/33). This is low but still well above the accuracy we found using data from the ten Brinke et al. study.

References

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review*, *10*, 214–234.

Bühlmann, P. (2004). Bagging, boosting and ensemble methods. In J. Gentle, W. Härdle, & Y. Mori (Eds.), *Handbook of computational statistics: Concepts and methods* (pp. 877–907). Berlin, Germany: Springer.

Eich, E. (2014). Business not as usual. *Psychological Science*, *25*, 3–6.

Ioannidis, J. P. A. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science*, *7*, 645–654.

Levine, T. R., & Bond, C. F. (2014). Direct and indirect measures of lie detection tell the same story: A reply to ten Brinke, Stimson, and Carney (2014). *Psychological Science*, *25*, 1960–1961.

Loftus, E. (2003). Our changeable memories: Legal and practical implications. *Nature Reviews Neuroscience*, *4*, 231–234.

MacKay, D. J. C. (2003). *Information theory, inference & learning algorithms*. New York, NY: Cambridge University Press.

McDowell, M. A., Fryar, C. D., Ogden, C. L., & Flegal, K. M. (2008). *Anthropometric reference data for children and adults: United States, 2003–2006* (National Health Statistics Reports No. 10). Retrieved from the National Center for Health Statistics Web site: http://www.cdc.gov/nchs/data/nhsr/nhsr010.pdf

R Development Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. New York, NY: Cambridge University Press.

ten Brinke, L., & Carney, D. R. (2014). Wanted: Direct comparisons of unconscious and conscious lie detection. *Psychological Science*, *25*, 1962–1963.

ten Brinke, L., Stimson, D., & Carney, D. R. (2014). Some evidence for unconscious lie detection. *Psychological Science*, *25*, 1098–1105.

Ulrich, R., & Miller, J. (1993). Information processing models generating lognormally distributed reaction times. *Journal of Mathematical Psychology*, *37*, 513–525.