Law and Human Behavior, Vol. 31, No. 1, February 2007 (© 2007)
DOI: 10.1007/s10979-006-9016-1

On Lie Detection "Wizards"

Charles F. Bond Jr.[1,3] **and Ahmet Uysal**[2]

Published online: 13 January 2007

M. O'Sullivan and P. Ekman (2004) claim to have discovered 29 wizards of deception detection. The present commentary offers a statistical critique of the evidence for this claim. Analyses reveal that chance can explain results that the authors attribute to wizardry. Thus, by the usual statistical logic of psychological research, O'Sullivan and Ekman's claims about wizardry are gratuitous. Even so, there may be individuals whose wizardry remains to be uncovered. Thus, the commentary outlines forms of evidence that are (and are not) capable of diagnosing lie detection wizardry.

KEY WORDS: deception detection.

Legal professionals have an interest in deception. Many of them believe that deception is common. Many of them have views about their own ability to detect deceit.

Psychologists have been writing about deception for decades. They base their writings on psychological research. Hundreds of relevant research studies have been conducted over the years, and these have been summarized in a recent review (Bond & DePaulo, in press). The accumulated research literature suggests that people are not good at detecting deception. In judging whether or not others are lying, the average person is accurate roughly 54% of the time when 50% would be expected by chance.

Deception researchers have studied topics of forensic interest. Some have documented the lie detection abilities of relevant occupational groups (Meissner & Kassin, 2002). Others have attempted to develop lie detection training procedures that would be of use in forensic settings (Kassin & Fong, 1999). Many have focused their research efforts on the sorts of high-stakes lies that would be of interest in the legal arena (Garrido, Masip, & Herrero, 2002). For book-long treatments of these and other forensic aspects of deception, see Vrij (2001) as well as Granhag and Stromwall (2004).

[1] Department of Psychology, Texas Christian University, Fort Worth, Texas.
[2] Middle Eastern Technical University, Ankara, Turkey.
[3] To whom correspondence should be addressed at Department of Psychology, P. O. Box 298920, Texas Christian University, Fort Worth, Texas 76129; e-mail: c.bond@tcu.edu.

O’Sullivan and Ekman (2004) have recently reported a curious result. They found 29 “wizards” of deception detection. Interviews reveal that these “geniuses” of lie detection have “an intense focus and investment in their performance” (O’Sullivan, 2005, p. 246).

Lie detection wizards should be of interest to forensic researchers. It would be smart for researchers to scrutinize wizards’ occupational backgrounds, to see if legal professionals are overrepresented. Those who are attempting to develop lie detection training procedures should wish to uncover wizards’ detection strategies. All forensic researchers should be interested in wizards who have demonstrated an ability to detect high-stakes lies, and this is precisely the ability claimed for these 29 wizards (O’Sullivan & Ekman, 2004). Some of the lies they detected were told under threats of punishment. On the more practical side, law enforcement agencies would be wise to hire lie detection wizards. The US government should also want to hire them. Recognizing this latter point, O’Sullivan and Ekman (2004) have already “suggested to government officials that wizards . . . might be consulted on cases of extraordinary importance” (p. 284).

Before concluding that these 29 individuals have extraordinary abilities, let us scrutinize the statistical evidence for wizardry at lie detection. In this commentary, we show that a simple chance mechanism could produce the wizard-like performances O’Sullivan and Ekman observed. We also describe the forms of evidence that would be needed to determine whether or not people are lie detection wizards.

“WIZARDS” OF LIE DETECTION

O’Sullivan and Ekman (2004) found 29 individuals who performed well on three high-stakes lie detection tests. Test 1 involves lies about opinions; test 2 involves lies about a mock crime; and test 3 involves lies about emotion. These tests are in a videotaped format. Ten people appear on each tape—five who are lying and five who are telling the truth. Test-takers see each person on each tape. In response to each videotape segment, test-takers indicate whether or not the person on the tape is lying.

Results on the opinion, crime, and emotion tests have appeared in seven journal articles (Ekman & O’Sullivan, 1991; Ekman, O’Sullivan, & Frank, 1999; Etcoff, Ekman, Magee, & Frank, 2000; Frank & Ekman, 1997; Frank, Paolantonio, Feeley, & Servoss, 2004; O’Sullivan, 2003; O’Sullivan, Ekman, & Friesen, 1988). Let us summarize the findings presented in those professional outlets while confining our attention to test results for college students. Accuracy rates on the opinion, crime, and the emotion test are available for 353, 113, and 464 undergraduates who judged those tapes. On the average, these students correctly judged 56.06, 60.19, and 50.00% of the opinion, crime, and emotion segments.

O’Sullivan and Ekman have screened over 12,000 professionals for their expertise at lie catching. Twenty-nine of the 12,000 met the researchers’ criteria for wizardry. To convince readers of the genius displayed by these individuals, O’Sullivan and Ekman report a statistical computation. It shows that there is an extremely low probability that any given person would achieve a “wizard” performance by chance.

For purposes of the computation, the researchers define a "chance" performance as one in which a test-taker responds to video segments as though s/he were flipping coins—thus having a 50/50 chance of being correct on each of 30 lie-or-truth segments.

We have several statistical concerns about this work. Let us begin by noting that people do not respond to the crime and opinion video segments as though they were flipping coins. They judge more than 50% of these lies and truths correctly, as journal articles report. Also, O'Sullivan and Ekman (2004) report the probability of *one* individual achieving a wizard-qualifying test score, even though more than 12,000 individuals have been tested. Although the probability of any one person excelling at these tests is low, it may nonetheless be quite likely that one (or more) *of 12,000* test-takers would achieve a high score (cf., Nickerson & Hammond, 1993). An appropriate model for these data would need to acknowledge the large number of people who have taken the tests.

Here we consider the possibility that none of the 12,000 test-takers were wizards. Under this null hypothesis, we compute the probability that 29 (or more) of the test-takers would meet the criteria O'Sullivan and Ekman have used to identify wizards. We determine this probability under two statistical models: a coin-flipping model and a research-based model. We regard the first model as inappropriate and are presenting it only because it formalizes an assumption made by O'Sullivan and Ekman. We regard the second model as more realistic.

Let us explain these two models. The coin-flipping model begins with an assumption that is false, according to earlier research—that people have a 50/50 chance of being accurate on each video segment of each test O'Sullivan and Ekman use to identify wizards. It assumes that 50% of nonwizards would judge the average lie accurately, when more than 50% have judged it accurately in prior research. Because the model assumes that fewer nonwizards would make correct judgments than the number who would actually make them, it should lead to an underestimate of the number of "wizards" on these tests. Anyone who gets substantially more than 50% right will be labeled a "wizard." But the average judge gets more than 50% right, as previous research shows.

The research-based model uses results from earlier student judges to estimate a chance level of accuracy on each segment. If (despite Meissner & Kassin, 2002), we entertain the possibility that students are less accurate on these lie detection tests than professionals, this model may also be conservative. Even so, the research-based model for predicting professionals' lie detection accuracy is more realistic than the coin-flipping model. Hence, it will be interesting to compare the number of "wizards" O'Sullivan and Ekman found with a research-based prediction for the number of wizards.

Let us now offer a few statistical details. By the O'Sullivan and Ekman (2004) definition, a person is a "wizard at deception detection" if the person

(a) reports correct judgments to at least nine opinion video segments, and gives correct judgments to *either*

(b) at least eight crime video segments *or*
(c) at least eight emotion video segments.

We begin with the simplifying assumption that all of the video segments on a given lie detection test are equally difficult. Given our simplifying assumption, we use the binomial distribution to determine the probability of a performance on each test that satisfies the cutoff score prescribed by the authors.

Let us symbolize the probability of such a performance as x for the opinion test, y for the crime test, and z for the emotion test. Then for reasons that are explained by Howell (2002), the overall probability of a person meeting the O'Sullivan and Ekman criterion for wizardry can be determined from the equation $x(y + z - yz)$. Let us symbolize this overall probability as p. If 12,000 randomly selected nonwizards took these tests, we would expect 12,000 × p of those individuals to be classified as wizards.

Let us now present some of the numerical results from two statistical models: the coin-flipping model and the research-based model. Under the coin-flipping model, $x = .0114$, $y = .0553$, and $z = .0553$. These figures imply that $p = .00122$. Thus, we would expect 14.71 of the 12,000 test-takers to meet the O'Sullivan and Ekman criterion for being a wizard. Under the research-based model, $x = .0270$, $y = .1701$, and $z = .0553$. This model implies that $p = .00583$. Thus if we use results from earlier student judges, we would expect 69.98 of the 12,000 individuals to meet the criterion for being a wizard.

O'Sullivan and Ekman tested 12,000 individuals for wizardry. They classified 29 of those 12,000 as wizards (95% confidence interval = 18.46–39.54 wizards). O'Sullivan and Ekman found more wizards than would be expected under the coin-flipping model. In our way of thinking, this discrepancy verifies what was already known—that people do not respond to these video segments as though they were flipping coins. The coin-flipping model classifies anyone who achieves substantially more than 50% accuracy on these deception tests as a wizard. But the average student achieves more than 50% accuracy. That is why more people are classified as "wizards" than the coin-flipping model predicted.

O'Sullivan and Ekman classified only 29 professionals as wizards. This is fewer than the 70 wizards our research-based model would seem to predict. This discrepancy is not easy to explain. One possible explanation relates to a detail of our research-based model—its use of only student judges to predict lie detection accuracy. In addition to the 930 students who judged the three relevant videotapes, 508 other people had judged these tapes before the wizards made any judgments. Suppose we constructed a research-based model from the accuracy of *all* 1438 judges of these videotapes. Although one might expect this more inclusive research-based model to yield more accurate predictions, it in fact predicts that 155 of the 12,000 professionals should be classified as wizards. Thus, it does nothing to explain why only 29 "wizards" were found.

Although we cannot be sure why the observed number of wizards is so much lower than our research-based prediction, let us venture two possible explanations: one is methodological, the other statistical. The methodological explanation is that these 12,000 professionals took lie detection tests under different conditions than earlier judges. For example, the professionals took two of the tests in their private residences; whereas, the earlier judges took those tests in an academic group setting. Perhaps the professionals were tested under conditions less conductive to lie detection.

The second, statistical explanation begins by noting that only a subset of the 12,000 professionals studied by O'Sullivan and Ekman took all three lie/truth discrimination tests. Let us elaborate on the researchers' testing protocol. The 12,000 professionals in this study attended group workshops on detecting deception. At the outset of each workshop, participants judged 10 opinion lie-or-truth segments. After completing this opinion test, participants were given the correct answers "and asked to report to the group (by a show of hands) how many items they got right" (O'Sullivan & Ekman, 2004). Workshop participants who reported scoring 90% or better were invited to take additional tests. Some accepted the invitation and subsequently completed the crime and emotion tests.

We do not know how many professionals completed the three-test lie detection battery, but if we are allowed a post hoc assumption about that number, we can fit the research-based model to these data perfectly. To do so, we note that under the research-based model there is a .00583 probability of a nonwizard being classified as a wizard. Thus, the assumption that we must make is that N people took all three lie detection tests, where N can be found from the equation $29 = .00583\ N$. Algebra then reveals that to "predict" 29 wizards, we must assume that 5,390 of the 12,000 professionals completed all three lie detection tests. Perhaps it is implausible to assume that 5,390 of 12,000 professionals completed all three of the researchers' lie detection tests. Readers need to make this judgment for themselves, after studying the article by O'Sullivan and Ekman (2004).[4]

DIAGNOSTIC EVIDENCE

As statisticians, we are in no position to conclude that the 29 individuals whom O'Sullivan regards as lie detection wizards are *not*, in fact, wizards. Following standard null hypothesis logic, failures to find significant evidence may represent Type II errors. Thus, the current evidence may simply be insufficient to uncover these 29 individuals' wizardry. Indeed, it is conceivable that many more of the 12,000 professionals in question are wizards, and the authors' procedures lack sufficient power to detect this fact. In light of this logical possibility, it is important to consider the forms of evidence that could (and could not) diagnose lie detection wizardry.

[4] It is under the model based on student judges' lie detection accuracy that we must make the assumption outlined in the text. Under the model based on the accuracy of all earlier judges, we "predict" 29 wizards by assuming that 2,248 of the 12,000 professionals completed all of the lie detection tests. An alternative statistical analysis would treat as unknown the number of professionals who were invited to take all of the lie detection tests. Then it would use the chance probability of wizard-qualifying performances on the latter two tests to induce the number of professionals who in fact completed the test battery. Applying this alternative treatment in the context of the student judge research-based model, we "predict" 29 wizards if 114 of the 12,000 professionals completed all three tests. Obviously, our two statistical analyses yield different inferences. We ourselves regard the second analysis as more defensible—given that participants were (or were not) invited to complete the test battery based on their self-reports of a test outcome.

To be diagnostic, tests for wizardry must meet certain psychometric and statistical criteria. The tests must be internally consistent. Because the tests must show high composite reliability, they may need to be long (Nunnally, 1978). An a priori statistical definition of wizardry needs to be specified. Computations must demonstrate that the cutoff score to be used for identifying wizards has sufficient statistical power to detect this statistically defined “wizard.” This cutoff score should take into account probabilistic arguments of the sort we have outlined above. Evidence of an extra-chance detection performance should be required before any wizardry is inferred. A second, lower cutoff score should also be developed for the identification of nonwizards.

Of course, there are other considerations in testing. If a lie detection test is to be diagnostic of a general forensic ability, it should expose test-takers to a variety of forensically relevant lies. There would be advantages if the person who developed this test was not the person who interpreted the test. If the test developer and interpreter had different theoretical orientations, it would be hard to contend that the content of the test and the interpretation of the test were related to one another in a way that reflected “bias” (Rosenthal, 1994).

Standardized testing procedures should be used. No one should be allowed to score their own test of lie detection ability, nor should anyone’s report of their score on a test be uncritically accepted. Tests should be supervised and scored by a third party, ideally an individual who does not know the correct answers to test items. The supervisor should have no preexisting belief about the ability level of any particular test-taker. The absence of any preexisting supervisor assumptions would be most critical if the testing format were interactive—say, if the test required the supervisor to interview subjects. If the intended interpretation is that a test measures a preexisting ability, test-takers should get no preexposure to the sorts of items that appear on the test—other than an orienting instructional item or two.

CONCLUSION

Convincing evidence of lie detection wizardry has never been presented. In fact, no truly diagnostic procedure for identifying wizards has ever been reported. Thus, it is premature for government officials to rely on any “wizards” of lie detection in cases of extraordinary importance. It is likewise premature for forensic psychologists to base their research on the assumption that lie detection wizards do (or do not) exist. Rather than constructing research programs on any such assumption, forensic psychologists would be better advised to spend their time developing procedures that could in principle distinguish lie detection wizards from nonwizards.

REFERENCES

Bond, C. F., Jr., & DePaulo, B. M. (in press). Accuracy of deception judgments. *Personality and Social Psychology Review*.

Ekman, P., & O'Sullivan, M. (1991). Who can catch a liar? *American Psychologist*, *46,* 913–920.

Ekman, P., O'Sullivan, M., & Frank, M. G. (1999). A few can catch a liar. *Psychological Science*, *10,* 263–266.

Etcoff, N., Ekman, P., Magee, J. J., & Frank, M. G. (2000). Lie detection and language comprehension. *Nature*, *405,* 139.

Frank, M. G., & Ekman, P. (1997). The ability to detect deception generalizes across different types of high-stakes lies. *Journal of Personality and Social Psychology*, *72,* 1429–1439.

Frank, M. G., Paolantonio, N., Feeley, T. H., & Servoss, T. J. (2004). Individual and small group accuracy in judging truthful and deceptive communication. *Group Decision and Negotiation*, *13,* 45–59.

Garrido, E., Masip, J., & Herrero, C. (2002). Police officers' credibility judgments: Accuracy and estimated ability. *International Journal of Psychology*, *39,* 276–289.

Granhag, P. A., & Stromwall, L. A. (2004). *Deception detection in forensic contexts*. Cambridge, England: Cambridge University Press.

Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Pacific Grove, CA: Duxbury Press.

Kassin, S. M., & Fong, C. T. (1999). "I'm innocent!": Effects of training on judgments of truth and deception in the interrogation room. *Law and Human Behavior*, *23,* 499–516.

Meissner, C. A., & Kassin, S. M. (2002). "He's guilty:" Investigator bias and judgments of truth and deception. *Law and Human Behavior*, *26,* 469–480.

Nickerson, C. A. E., & Hammond, K. R. (1993). Comment on Ekman and O'Sullivan. *American Psychologist*, *48*, 989.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw Hill.

O'Sullivan, M. (2003). The fundamental attribution error in detecting deception: The boy-who-cried-wolf effect. *Personality and Social Psychology Bulletin*, *29,* 1316–1327.

O'Sullivan, M. (2005). Emotional intelligence and deception detection: Why most people can't 'read' others, but a few can. In R. E. Riggio & R. S. Feldman (Eds.), *Applications of nonverbal communication* (pp. 215–253). Mahwah, NJ: Erlbaum.

O'Sullivan, M., & Ekman, P. (2004). The wizards of deception detection. In P. A. Granhag & L. A. Stromwall (Eds.), *Deception detection in forensic contexts* (pp. 269–286). Cambridge, UK: Cambridge Press.

O'Sullivan, M., Ekman, P., & Friesen, W. V. (1988). The effect of comparisons on detecting deceit. *Journal of Nonverbal Behavior*.

Rosenthal, R. (1994). On being one's own case study: Experimenter effects in behavioral research—30 years later. In W. Shadish & S. Fuller (Eds.), *The social psychology of science* (pp. 214–229). New York: Guilford Press.

Vrij, A. (2001). *Detecting lies and deceit: The psychology of lying and the implications for professional practice*. New York: Wiley.