Law Hum Behav (2007) 31:117–123
DOI 10.1007/s10979-006-9058-4

ORIGINAL ARTICLE

Unicorns or Tiger Woods: Are Lie Detection Experts Myths or Rarities? A Response to *On Lie Detection "Wizards"* by Bond and Uysal

Maureen O'Sullivan

Published online: 13 January 2007

Abstract Bond and Uysal (this issue) complain that expert lie detectors identified by O'Sullivan and Ekman (2004) are statistical flukes. They ignore one class of experts we have identified and misrepresent the procedures we use to identify the others. They also question the psychometric validity of the measures and protocol used. Many of their points are addressed in the chapter they criticize. The fruitfulness of the O'Sullivan-Ekman protocol is illustrated with respect to improved identification of expert lie detectors, as well as a replicated pattern of errors made by experts from different professional groups. The statistical arguments offered confuse the theoretical use of the binomial with the empirical use of the normal distribution. Data are provided that may clarify this distinction

Keywords Deception · Lie detection · Accuracy · Expertise

Bond and Uysal criticize a chapter (O'Sullivan & Ekman, 2004) describing the beginning stages of an ongoing research program, written to share early results with others interested in similar issues. They raise two objections: 1) The expert lie detectors are not really expert since their scores could have occurred by chance alone. 2) The testing protocol used does not meet the requirements of classical psychometric test theory. These issues were addressed, in whole or in part, in the chapter under consideration, but since their treatment was either overlooked or unconvincing, unto the fray let us repair.

Bond and Uysal (this issue) suggest a formula to estimate the probability of obtaining at least 90% on one test and at least 80% on one of two additional tests. This formula is misleading and inaccurate in two regards: (1) At the time the 2004 chapter was written, we had identified 29 expert lie detectors of whom 14 obtained scores of at least 90% on the first test and at least 80% on **both** of the succeeding two tests. As reported in the 2004 chapter, the odds of this occurring by chance alone is $p(x)p(y)p(z)$. With a mean of 50%, this value is (.01)(.05)(.05) or .000025. Even if one assumes a mean of 54% for each of the three tests [the average lie detection

M. O'Sullivan (✉)
Department of Psychology, University of San Francisco,
2130 Fulton Street, San Francisco, CA 94117
e-mail: osullivan@usfca.edu

accuracy Bond and DePaulo (in press) report in a recent meta analysis] the probability of such an occurrence by chance alone is .00016 [(.02)(.09)(.09)].

2) In the 2004 chapter, we clearly indicated that our protocol is sequential. Participants "qualify" for consideration as expert lie detectors by first obtaining a score of 90% on the opinion test and only these individuals are given additional tests. Bond and Uysal calculated their formula as though all participants received all measures. It is more appropriate to apply their formula only to the much smaller number who completed all three tests. The largest single group we tested consisted of 1057 therapists. Forty eight of them obtained scores of 90% on the "qualifying" opinion test. Of these 48, to date, 22 have completed all three tests. The Bond-Uysal formula suggests that .00583 or .128 (less than one) of these 22 people should obtain scores of 90% on one test and 80% or better on one of the other two. In actuality, six such experts were identified. Even more interesting, however, is which tests these therapists did well on. As reported by B & U, the average score for one group of students on the crime test was 60%. By chance only, with a mean of 60%, one would expect 17% or four of the 22 therapists to obtain scores of 80% or better on this test, but only one of them did. On the other hand, with the emotion test which has a student mean of 50%, one would expect 5% or only one of the 22 therapists to achieve such a score, yet five of the therapists did so. When one obtains a finding that occurs five times more often than would be predicted by chance alone, chance is not the most parsimonious explanation for the phenomenon.

As B & U note, the means obtained by some professional groups, such as those containing a high proportion of expert lie, detectors are even higher (i.e., 68%, 73% see Ekman, O'Sullivan & Frank, 1999) than those obtained with students, further undermining the argument that the distribution of lie detection scores, at least in certain groups, reflects chance alone. Yet Bond and Uysal, while agreeing that lie detection is not chance, use a chance distribution to examine their thesis. A more appropriate distribution is the normal curve. Although the binomial and the normal curve are similar, they proceed from different assumptions and are the bases for different kinds of statements. Probability distributions (like the binomial) allow one to state "Given 12,000 golfers, one of whom is Tiger Woods, the chance of selecting Tiger Woods from among the 12,000 golfers, *by chance alone*, is one in twelve thousand." The normal distribution, based on empirical data from golf tournaments, expert ratings and the like, allows one to conclude: "Tiger Woods is among the top one percent of 12,000 golfers tested." Our model of expert lie detection is the latter, not the former.

The normal distribution has been found to underlie the empirical frequency distributions of many psychological variables, such as intelligence (Wechsler, 1958) and personality (Costa & McCrae, 1985). This is the model from which the expert lie detection research proceeded, based on three assumptions: 1) Lie detection is an ability that can be measured; 2) This ability is distributed like many other abilities (i.e., normally); 3) **Therefore, only a very few people will be highly accurate.** In research on human intelligence, those individuals who score two standard deviations above the mean are sometimes termed "geniuses." In a large, unselected sample, fewer than two percent of people would be so classified. If a large group has a mean IQ of 100 with a standard deviation of 15, about two percent of that group would be expected to achieve IQ scores of 130 or above. If the mean of the group is 130, then, obviously, 50% of the group will achieve scores of 130 or above. This is not a bias in the test, but a difference in the ability level of the two groups. It is misleading to argue both that the lie detection tests we used generate scores based only on chance and then to argue that lie detection is not a chance phenomenon and therefore a "research based" mean of 59% should be used. In a large sample, a mean of 59% would be significantly different from chance. Using this significantly-different-than-chance mean, Bond and Uysal then suggest that not only are our expert lie detectors statistical flukes but that we should have found more of them.

Why are there not more wizards? This issue was addressed in the 2004 chapter.

> Our task limits the lie detector to watching a one-minute videotape of a liar or truth teller being interviewed by someone else. The sample of behavior is quite limited (one minute), the observer does not have the opportunity to obtain a baseline sample of behavior that is known to be honest to use as a comparison, the observer is limited to merely watching the behavior on the videotape. Highly skilled interviewers, who have learned over the years, how to use their appearance and personality to best effect, or how to interview different personality types with different techniques, or who have developed particularly effective interview strategies for use in detecting deception may be handicapped by the protocol we use which does not allow them to use these talents and techniques. On the other hand, in earlier research . . . interviewers identified by their agencies as superior interviewers also were identified as superior lie detectors by . . . the measures used in the present study. **Although not all kinds of expert lie detectors will be detected by our methods, many of them will be.** (bold added) (O'Sullivan & Ekman, pp. 278–279)

All psychological measures are limited, as is ours. We have not identified all the experts in the samples we surveyed, only some of them. Potential experts had to be motivated to participate, to complete the follow-up tests, to meet with us and allow us to interview them. Because of the sensitivity of their jobs, some experts did not wish to be interviewed. These factors restrict the size of the final sample and may bias its constitution. But that does not mean that our sample does not include highly accurate lie detectors, only that they are lie detectors identified in a particular way. Certainly there are other expert lie detectors, including those adept at recognizing kinds of lies not examined in the measures we used. Although our procedure is subject to false negatives, given the low scores achieved by most other subjects and the requirement that experts do well on at least two of three tests, susceptibility to false positives among wizards who do well on two of the three lie detection tests is low. For wizards identified on the basis of high scores on **all three tests** (14 of the 29 discussed in the 2004 chapter), the probability of their classification being due to chance alone is extremely low (one in 25 million).

What other evidence suggests that the expert lie detectors are really expert?

Four additional sources of evidence, difficult to dismiss on the basis of statistical chance alone, can be offered: 1) the kinds of errors different types of wizards make; 2) the differential occurrence of wizards in different occupational groups; 3) the ability to "predict" wizards; and 4) the differential responses of wizards and non-wizards.

1. *Error Patterns*. As noted above, six therapists obtained scores of 80% on one of the two additional tests they were given. Of these, five were highly accurate on the emotion lie task (nurses watching nature or surgical films). Only one therapist was accurate on the crime task (men lying or telling the truth about whether they had stolen money) and not the emotion lie task. (A seventh therapist from this study was highly accurate on all three tasks.) Chance alone does not predict that five of six therapists will do well on the emotion test, but not the crime test, especially considering that average accuracy is higher for the crime test.
   Another way to look at this finding was reported in the 2004 chapter. We described three types of wizards (pp. 275–277): 1) those who scored 80% or better on all three tests ($n = 14$) (the group that Bond and Uysal ignore); 2) those who scored 80% or better on the opinion and the emotion tests, but not on the crime test ($n = 7$ including the six from the therapist study discussed earlier) and 3) those who scored 80% or better on the opinion and the crime test, but not on the emotion test ($n = 8$). All of the experts who showed pattern 2 were therapists. All but one of the experts who showed pattern 3 were law-related professionals. A chi square analysis of these data (O'Sullivan & Ekman, 2004, p. 276) was highly significant. This error pattern has been replicated with additional wizards who have been identified in the last three years. If the wizards are mere chance occurrences, it is difficult to explain this plausible profession-related finding, and its replication.

2. *Group Yield Rates*. Forthcoming publications will describe the score distributions of a wide variety of professional and other groups and the relative incidence of wizard occurrence in each of them. These data will both replicate data already in the literature and provide new insights into the factors that contribute to accurate lie detection. For example, it is already known that unselected police officers, as a group, score close to chance in their lie detecting accuracy, but that selected groups such as Secret Service agents (Ekman & O'Sullivan, 1991) or law enforcement personnel interested in and motivated to improve their lie detecting ability (Ekman, O'Sullivan & Frank, 1999) are significantly above average. (These expert groups were identified before the start of the Expert Lie Detector project. It was the discovery of these highly accurate groups that led us to believe that, with sufficient testing, a large enough sample of highly expert individuals could be identified.) We have replicated these findings with other law enforcement groups several times, thereby undermining the hypothesis of a chance occurrence. Obviously, a response to a commentary does not provide sufficient scope to report these replications as well as the new data gathered on other professional groups.
3. *Predicting Expert Lie Detectors*. The search for accurate human lie detectors identified with objectively-scored group administered tests has gone on at least since Fay and Middleton's 1941 study. Despite this longstanding effort, few highly accurate groups of lie detectors have been identified. When the 2004 chapter was written, 29 expert lie detectors had been identified. Since then, 13 more experts have been identified. It took almost 30 years (Ekman & Friesen, 1974) to find 29 expert lie detectors, but only three years to find an additional 13, because we learned which groups were more likely to yield wizards and focused our testing efforts on them.
4. *The proof of the pudding: What wizards see, hear and think. Who wizards are*. Interviews with, and objective assessments of, the expert lie detectors will not be completed until the end of 2006. Their responses will be compared with those of controls similar in age, education, geographic location and social class who are not highly accurate in their lie detection abilities. To date, our analyses suggest differences in the kinds and numbers of clues observed as well as in the inferential processes involved in understanding others. Our analyses have suggested new clues to lie detection and methods for refining the analyses of clues already identified in the literature. This knowledge can be used to enhance automated lie detection methods, guide neurological examination of the process of lie detection, improve the conduct of interviews used to detection deception and contribute to the training and selection of personnel involved in interviewing and judging others. This addition to knowledge is another, albeit indirect, validation of their classification as expert lie detectors.

   Another indirect validation is the kinds of people the truth wizards are. They are not ordinary. Most are of their professions—federal agents chosen to be profilers, to work on the Unabomber case, to head the Secret Service White House detail; therapists who have written books on expertise in counseling; nationally known arbitrators; investigators singled out for media attention because of their success in closing cases or to head training programs because of their interviewing abilities.

Psychometric concerns

Bond and Uysal offer a variety of observations about what a "good" expert lie detection test should look like, implying that our tests and our protocol do not meet these criteria. They proceed from the perspective of classical test theory (Nunnally, 1978; Gulliksen, 1950) in which psychological tests are conceptualized as relatively static composites of similar items, each of which contributes an additional, but essentially similar, datum to the total score. They suggest

that the most desirable type of reliability is item homogeneity, usually measured by Cronbach's alpha. But recently, Cronbach and Shavelson (2004) have argued that item homogeneity is not the only or even the most desirable estimate of reliability in all cases.

Another recent development in psychometric theory is item response theory (Embretson & Reise, 2000), which proceeds from the assumption that different items will be differentially diagnostic for different ability levels. IRT also suggests that tests be carefully assessed at the item level. The tasks we used contain many clues related to detecting deception, including but not limited to behavioral clues in many channels (Ekman, 2001; DePaulo et al., 2003), interactional clues (Burgoon, Buller, White, Afifi, & Buslig, 1999), personality clues, social class and interest clues, gender, ethnicity, etc. In addition, accurate lie detection of a given individual necessitates understanding what kind of person that individual is (O'Sullivan, 2003, 2005), so the abilities involved in accurate impression formation (Ickes, 1993; Funder, 1999) are also recruited. Research with the lie detection measures used to identify the "truth wizards" suggests that each item involves different kinds of abilities. Some abilities, such as those involved in recognizing fleeting or subtle emotions (Ekman & O'Sullivan, 1991; Frank & Ekman, 1997) occur in many items; other skills are involved in only a few items. Hence, homogeneity is not a goal. Rather, a careful analysis of the relationship between each item and "wizards" of different types (as described above) is more likely to yield useful information about how to detect different kinds of lies accurately.

Psychometricians have long argued that construct validation is an ongoing *process* within which the meaning of test scores can be systematically understood and that although convergent and discriminant validity remain important, the ultimate value of a test lies in what it permits one to discover (Messick, 1995). The convergent validity of these tests was addressed by Frank and Ekman (1997) who demonstrated a high concurrence in two studies between people classified as accurate and inaccurate on the opinion and crime measures (χ^2 (1) = 6.15, $p < .02$, p. 1434; $\chi^2(1) = 5.46$, $p < .025$, p. 1436). Additional convergent validity was provided by significant positive correlations, as predicted, between the ability to recognize rapidly occurring facial expressions of emotion and the crime test ($r = .34$, $p < .04$) (Frank and Ekman, 1997, p. 1436) as well as the emotion test ($r = .270$, $p = .02$) (Ekman & O'Sullivan, 1991, p. 917). Also as predicted, accurate as opposed to inaccurate observers of truthfulness were significantly more likely to report attending to nonverbal clues (χ^2 (2) = 45.5, $p < .001$; $\chi^2(2) = 10.96$, $p < .01$) (Ekman & O'Sullivan, 1991, p. 918).

Other indications of convergent validity are contained in the information provided earlier, in which certain professional groups are found to contain more wizards than others, i.e., certain types of professional experience seem to be related to certain kinds of lie detection accuracy. We have made a series of predictions about differences in the perceptual and cognitive capacities of wizards and their matched controls. If these differences are found, this will provide further evidence of convergent validity.

At the same time, the intercorrelations among the three lie detection tests, while significant, are not substantial. This finding is consistent with the view that different kinds of lies may involve different kinds of clues or may involve different decoding abilities. A very few expert lie detectors (Type 1) will be accurate at all or most kinds of lies; others (Types 2 and 3) will be accurate at lies with which they are familiar based on their life or professional experience, or because of their particular lie detection strategies. We have examined accuracy with three different kinds of high stakes lies. Obviously, the generalization of accuracy with these tests to accuracy with additional kinds of lies, as such measures become available, will also contribute to the demonstration of convergent validity.

Discriminant validity is being addressed through the use of contrasted groups in which the responses of accurate and inaccurate lie detectors are compared. This approach is widely used

in research on expertise, such as studies of grand masters in chess (DeGroot, 1946/1978), internationally known violinists, and outstanding medical diagnosticians (Ericsson, 2005; Ericsson & Simon, 1998; Ericsson & Charness, 1994) as well as in validating clinical tests such as the MMPI. As in these studies, the differences found provide both a retroactive validation of the classification as well as new information about the psychological processes involved in being an accurate lie detector.

Bond and Uysal complain about other psychometric issues, which either misrepresent or ignore information contained in the chapter they are critiquing or are erroneous, such as the suggestion that tests should be scored by ". . . individuals who do not know the correct answers" (Bond and Uysal, this issue).

Summary

Statistical flukes occur, but they are rare and epiphenomenal, as are unicorns. Geniuses and virtuosos of various sorts occur. They are rare, but real, as is Tiger Woods, a genius or wizard on the golf course. All psychological measures and all research paradigms are limited, as are ours. The bulk of the evidence, however, suggests that there are individuals who are highly accurate in understanding other people, in knowing whether they are lying or telling the truth. How they do this, what clues they use, how they utilize those clues, how they finalize their judgments, are among the questions we hope this research program will allow us to answer.

Acknowledgments Many thanks to Paul Ekman, Clark Freshman, Dana Carney, Shirley McGuire, Ben Lewis, Michael Davis-Wilson, Susan Heidenreich, David Howell, and Paul Zeitz for their comments on this response and the issues raised by it.

References

Bond, C. F. Jr., & Uysal, A. (2007). On lie detection "wizards." *Law and Human Behavior*, *31*(1).
Bond, C. F. Jr., & Atoum, A. O. (2000). International deception. *Personality and Social Psychology Bulletin, 26*(3), 385–395.
Bond, C. F. Jr., & DePaulo, B. M. (in press). Accuracy of deception judgments. *Personality and Social Psychology Review*.
Burgoon, J. K., Buller, D. B., White, C. H., Afifi, W., & Buslig, A. L. S. (1999). The role of conversational involvement in deceptive interpersonal interactions. *Personality and Social Psychology Bulletin*, *25*(6), 669–685.
Costa, P. T., & McCrae, R. R. (1985). *The NEO Personality Inventory Manual*. Odessa, FL: Psychological Assessment Resources.
Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, *64*(3), 391–418.
De Groot, A. (1946/1978). *Thought and choice in chess*. The Hague: Mouton (Original work published 1946).
DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin*, *129*(1), 74–118.
Ekman, P. (2001). *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage* (3rd ed.). New York: W. W. Norton.
Ekman, P., & Friesen, W. F. (1974). Detecting deception from the body or face. *Journal of Personality & Social Psychology*, *29*(3), 288–298.
Ekman, P., & O'Sullivan, M. (1991). Who can catch a liar? *American Psychologist*, *46*(9), 913–920.
Ekman, P., O'Sullivan, M., & Frank, M. G. (1999). A few can catch a liar. *Psychological Science*, *10*(3), 263–266.
Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Erlbaum.
Ericsson, K. A. (2005). Recent advances in expertise research: A commentary on the contributions to the special issue. *Applied Cognitive Psychology*, *19*, 233–241.

Ericsson, K. A., & Charness, N. (1994). Expert performance: Its structure and acquisition. *American Psychologist*, *49*, 725–747.
Ericsson, K. A., & Simon, H. A. (1998). How to study thinking in everyday life: contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, culture and activity*, *5*(3), 178–186.
Frank, M. G., & Ekman, P. (1997). The ability to detect deceit generalizes across different types of high-stake lies. *Journal of Personality and Social Psychology*, *72*(6), 1429–1439.
Funder, D. (1999). *Personality Judgment: A Realistic Approach to Person Perception*. San Diego: Academic Press.
Gulliksen, H. (1950). *Theory of Mental* Tests. Oxford, UK: Wiley.
Ickes, W. (1993). Empathic accuracy. *Journal of Personality*, *61*, 587–610.
Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749.
Nunnally, J. C. (1978). *Introduction to Psychological Measurement* (2nd ed). New York: McGraw Hill.
O'Sullivan, M. (2003). The fundamental attribution error in detecting deceit: The boy-who-cried-wolf effect. *Personality and Social Psychology Bulletin*, *29*(10), 1316–1327.
O'Sullivan, M. (2005). Emotional intelligence and detecting deception. Why most people can't "read" others, but a few can. In Riggio, R. and Feldman, R. (Eds.), *Applications of Nonverbal Communication* (pp. 215–253). Mahway, NJ: Erlbaum.
O'Sullivan, M., & Ekman, P. (2004). The wizards of deception detection. In Granhag, P.A., & Strömwell, L. (Eds.), *The Detection of Deception in Forensic Contexts* (pp. 269–286). Cambridge, UK: Cambridge University Press.
Wechsler, D. (1958). *The Measurement of Adult Intelligence* (4th ed.). Oxford, UK: Williams and Wilkins.