RESEARCH ARTICLE

Evaluation of the effectiveness of a brief deception detection training program

Stephen Porter*, Marcus Juodis, Leanne M. ten Brinke, Raymond Klein and Kevin Wilson

aDepartment of Psychology, University of British Columbia Okanagan, British Columbia, Canada; bDepartment of Psychology, Dalhousie University, Nova Scotia, Canada

(Received 6 March 2009; final version received 29 May 2009)

The discrimination of genuine and falsified emotional displays is critical in many contexts, including healthcare, forensic, and airport security settings. Previous research has demonstrated that comprehensive (two-day) empirically based deception detection training can lead to moderate gains in judgment accuracy. However, for many professional groups such extensive training is not feasible due to time and resource limitations. In the present study, we evaluated the effectiveness of brief (three-hour) training. N = 26 (13 females, 13 males) healthcare professionals with experience in evaluating the validity of medical claims participated in the training workshop. Their performance on two deception detection tasks was measured pre- and post-training; the participants attempted to discriminate: (1) videotaped truthful and fabricated stories concerning emotional events and (2) sincere and falsified emotional facial expressions. Results indicated that participants’ overall accuracy on both tasks increased modestly from chance (M = 51.2%) to significantly above chance (M = 60.7%), primarily due to an increase in their improved ‘hit’ rate from pre- to post-training.

Keywords: deception; training; malingering

High-stakes emotional deception is common in healthcare settings, forensic contexts, and airport security settings (see Porter & ten Brinke, 2009a). For example, in a study of 33,531 neuropsychological referrals, it was estimated that malingering had occurred in 29% of personal injuries, 30% of disability claims, 19% of cases involving criminal matters, and 8% of other medical matters (Mittenberg, Patton, Canyock, & Condit, 2002). In such settings, deceivers must tell false emotional stories and falsify convincing emotional displays to support the fraudulent story. Similarly, during police investigations and in the courtroom, false emotional displays can be used to feign...
distress or remorse. The Supreme Court of Canada (1993) concluded that judges and jurors must carefully observe a witness to ‘adequately evaluate body language, facial expressions and other indicators of credibility that are not apparent from a written transcript’. Despite this, recent surveys of legal decision-makers suggest that they lack training in credibility assessment, hold major misconceptions about deceptive behavior, and hold false stereotypes about deceivers (Porter & ten Brinke, 2009b; Stromwall & Granhag, 2003; Vrij, Akehurst, & Knight, 2006). In airports, potential terrorists must conceal and falsify their emotions in order to realize their violent plans. The 9/11 terrorists, for example, may not have been successful in their task if not for their ability to conceal their intentions and emotions before and during boarding. As a result of the ensuing catastrophe, security officials developed a massive training initiative aimed at helping staff identify threats by recognizing concealed emotions among passengers (New York Times, 17 August 2006; USA Today, 25 September 2007). However, the validity of such training is unclear and most human judges are poor lie detectors, performing around chance in deception detection tasks (e.g., Bond & DePaulo, 2006; Vrij, 2008).

Given the consequences of the failed identification of false emotional displays in high-stakes contexts, the development and evaluation of deception detection training is essential. Although only a few studies have evaluated the validity of such training, knowledge regarding deceptive behavior has increased considerably in the past two decades (see DePaulo et al., 2003; Porter & ten Brinke, 2009a). An intensive, empirically based two-day training program evaluated a decade ago (Porter, Woodworth, & Birt, 2000) was modestly successful in increasing the deception detection accuracy among a group of federal parole officers from below chance to 76.7%. This training consisted of myth dissolution, the provision of empirically based knowledge about cues to deception, practice, feedback, and information review/knowledge testing. Participants also were encouraged to examine critically their own decision-making via discussions with colleagues in order to reduce ‘tunnel vision’ decision-making (Porter, McCabe, Woodworth, & Peace, 2007). More recently, researchers have observed the potential biasing effect of training, suggesting that programs aimed at lie detection or truth detection can create lie or truth biases, respectively, in post-training performance (Masip, Alonso, Garrido, & Herrero, 2008; also see Kassin & Fong, 1999). Masip et al. (2008) suggested that future training programs would benefit from the inclusion of behavioral cues to both deception and honesty in order to increase credibility discrimination – enhancing hit rates (correct identifications of deception) while simultaneously reducing false alarms (attributing deception to a genuine account). Further, we have become aware that a comprehensive two-day training workshop on deception is not viewed as feasible (in terms of time and resources) for many relevant professional groups; as such, we
decided to examine the validity of a considerably abbreviated and focused training session.

The purpose of this study was to evaluate the effectiveness of a brief deception detection training program for healthcare employees who evaluate the credibility of medical claims, including ‘psychological’ claims involving subjective trauma and distress that often cannot be substantiated by physical evidence. Training consisted of myth dissolution, information provision (the latest scientific information on cues to deception [with a focus on real and false facial expressions and verbal and non-verbal cues to deception], honesty, and medical malingering), practice, and feedback. The study used a within-subjects design and compared the performance of the professionals on two deception detection tasks (evaluating narratives and emotional expressions) from pre-training to post-training. Signal detection analyses were conducted to evaluate the impact of training on accuracy, discrimination, and response bias.

Method

Participants

$N = 26$ professionals (13 females, 13 males) in Canada, whose work directly or indirectly involved the evaluation of medical claims, voluntarily participated in the credibility training program. They included: nine lawyers, five case managers, three physicians, three insurance officers, four appeal judges, and two adjudicators. The number of years of their experience in the healthcare field ranged from four months to 29 years, with a mean of 11.34 years ($SD = 6.98$ years). None of the individuals had received any previous training in credibility assessment.

Materials

Deception detection training program. The training program was a half-day (three-hour) workshop led by the first author. The workshop followed the general (but updated) approach outlined in Porter et al. (2000) and involved three main components: (1) myth dissolution; (2) information provision; and (3) practice/feedback exercises. Myth dissolution involved the debunking of erroneous beliefs surrounding deception detection (e.g., the validity of a single ‘rule of thumb’ for detecting lies, that intuition should be used as a reliable/valid guide, etc.). It also covered myths specific to injury/illness-related insurance cases, including information on base rates and common stereotypes. The second component of the program involved information provision. Trainees were presented with a review of the latest scientific knowledge on the detection of deception in applied settings, with a focus on malingering, verbal cues/statement analysis, nonverbal cues, and facial expressions. To enhance the differentiation of honest from deceptive
emotional expressions, the program gave examples of genuine universal emotional expressions via a brief description of and facial actions associated with each (Ekman & Friesen, 1975; Ekman, Friesen, & Hagar, 2002) in addition to examples of corresponding false expressions and discussion of absent muscle actions. Finally, through practice and feedback participants were provided with an opportunity to learn from their successes and failures. Each participant completed a baseline pre-training test and a post-training test (discussed below) of credibility assessment ability. After each test was completed, the workshop leader provided correct answers to the group and highlighted cues to deception and truthfulness present in each video clip.

Pre-training and post-training tests. All participants completed the same pre-training and post-training tests; each comprised four videotaped personal injury narratives (two genuine, two deceptive), three videotaped negative life event descriptions, and eight videotaped facial expressions of happiness (four genuine, four fabricated). Personal injury narratives were created by videotaping volunteers describing one genuine and one falsified serious injury that they had or had supposedly experienced. Truthful and deceptive life event narratives were obtained from six volunteer incarcerated federal offenders who were recruited as part of a larger study examining deception in offender and non-offender populations (Porter et al., 2008). The offenders and their respective video clips were randomly selected from a pool of 27 offenders who gave permission to use their narratives for deception detection training purposes. Finally, the 16 facial expressions of happiness were created by videotaping volunteers who provided two genuine and two deceptive displays of happiness, evoked by responding to randomized presentations of emotional images chosen from the International Affective Picture System (IAPS: Lang, Bradley, & Cuthbert, 1999). Volunteers were asked to respond to two highly positive/happy and two highly negative images (e.g., distressed baby in an incubator), each with an expression of happiness. These 30 video clips were randomly, but evenly, distributed to be part of either the pre-training test or post-training test. Thus, each of the pre-test and post-test included 15 randomly selected video clips consisting of three negative life event stories by different offenders, four personal injury stories by two different people (each person relating one truthful and one deceptive account), and eight happiness expressions by two different volunteers (each person displaying two genuine and two false smiles).

Procedure
The training (and pre- and post-tests) occurred in a group format. A large auditorium projection screen with a sound system was used to present the training and test stimuli. For the pre-test, practice videos, and post-test, participants were informed of the nature of the three types of tasks and told
that some of the stimuli were truthful and some deceptive. Participants indicated whether in their opinion each story or facial expression clip was genuine or deceptive, with a 30-second break between the clips. Each clip was shown once and was presented in real-time. Following the pre-training test, feedback was provided to the participants along with discussion of the cues to deception or honesty exhibited in each video clip. Then, the main training components were delivered, and the post-training test occurred in the same manner as the pre-test and, again, was followed by feedback and discussion.

Results

Total accuracy

A 2 (Test: pre- vs post-training) X 2 (Task: emotional facial expressions vs narratives) repeated measures factorial analysis of variance (ANOVA) was conducted with each participant’s total accuracy score (%) entered as the dependent variable. The analysis indicated a significant main effect for test administered, $F(1, 25) = 7.32, p < .05, \eta^2_p = .23$. Participants obtained higher total accuracy scores on the post-test ($M = 60.7\%$ [95% CI = 54.1–66.7%] ± 2.8%) than the pre-test ($M = 51.2\%$ [95% CI = 45.7–56.7%] ± 2.7%). The ANOVA also showed a main effect for task type that approached significance, $F(1, 25) = 3.99, p = .057, \eta^2_p = .14$. Participants obtained slightly higher total accuracy scores on the facial expressions task ($M = 60\%$ [95% CI = 54.1–65.3%] ± 2.7%) than the narratives task ($M = 52.2\%$ [95% CI = 46.3–58.1%] ± 2.9%). The interaction, however, was not significant, $F(1, 25) = .612, p > .05, \eta^2_p = .24$.

The participants’ total accuracy scores on the facial expressions task increased from chance on the pre-test ($M = 56.3\%$ [95% CI = 49.1–63.4%] ± 3.5%) to significantly above chance on the post-test ($M = 63.1\%$ [95% CI = 55.1–71.1%] ± 3.9%). A one-tailed paired-samples $t$-test showed that the mean difference approached significance, $t(25) = 1.40, p = .09$. Similarly, accuracy scores on the narratives task increased from chance on the pre-test ($M = 46.2\%$ [95% CI = 38.8–53.5%] ± 3.6%) to significantly above chance on the post-test ($M = 58.2\%$ [95% CI = 50.3–66.2%] ± 3.9%). A one-tailed paired-samples $t$-test showed that the difference between these two means was significant, $t(25) = 2.55, p < .01$.

Signal detection analyses

Hit and false alarm rates. Hits or true positives were cases in which deceptive communication was correctly identified as deceptive. False alarms or false positives were cases in which truthful communication was misidentified as deception. Hit rates were calculated by dividing each of
their total number of hits by the total number of deceptive cases. False alarm rates were calculated for each participant by dividing their total number of false alarms by the total number of truthful cases. Table 1 summarizes the hit and false alarm rates in terms of means, 95% confidence intervals, and standard errors of the means for the facial expressions and narratives tasks for pre- and post-testing. The table shows noteworthy increases in participants’ hit rates but little change in their false alarm rates from pre- to post-testing.

**Discrimination accuracy (d’) and response bias (β)**

Whereas hit and false alarm rates are not independent, d’ and β are theoretically independent. d’ captures a participant’s discrimination accuracy, commonly referred to as ‘sensitivity’ (i.e., the ability to distinguish between truthfulness and deception). When d’ is positive and high, the participant is displaying good discrimination accuracy. When d’ is near zero, the participant is not discriminating between deceptive and truthful stimuli. Finally, if d’ is negative and low, the participant is misperceiving the stimuli such that deceptive behavior is labeled as truthful, and truthful behavior is labeled as deceptive. β provides a measure of a participant’s response bias, and captures the degree of evidence required for a participant to decide that deceptive communication is being observed. A positive β value indicates that a participant has a liberal response bias or a tendency to decide that deceptive behavior is being observed. A negative value indicates that a participant has a conservative response bias or a tendency to conclude that deceptive behavior is not being observed. When β is near zero, then the participant has no response bias and the two kinds of errors (misses and false alarms) are approximately equated.

A 2 (Test: pre-test vs post-test) X 2 (Task: emotional expressions vs narratives) repeated measures ANOVA was conducted on the d’ scores. There was a significant main effect of the test administered, $F(1, 25) = 15.44, p < .01, \eta^2_p = .38$ with mean d’ values increasing from

<table>
<thead>
<tr>
<th>Table 1. Means, 95% confidence intervals, and standard errors of the means for participants’ hit rate and false alarm rate for emotional facial expressions and narratives from pre-test to post-test.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type of task</strong></td>
</tr>
<tr>
<td>Emotional facial expressions</td>
</tr>
<tr>
<td>hit rate</td>
</tr>
<tr>
<td>false alarm rate</td>
</tr>
<tr>
<td>Narratives</td>
</tr>
<tr>
<td>hit rate</td>
</tr>
<tr>
<td>false alarm rate</td>
</tr>
</tbody>
</table>
pre- \((M = -0.10 \ [95\% \ CI = -0.372–0.16] \pm 0.13)\) to post-training \((M = 0.67 \ [95\% \ CI = 0.33–1.00] \pm 0.16)\). There was also a significant main effect of task, \(F(1, 25) = 6.88, p < .05, \eta_p^2 = .22\), with higher mean values of \(d'\) on the emotional facial expressions task \((M = 0.50 \ [95\% \ CI = 0.22–0.78] \pm 0.14)\) than on the narratives task \((M = 0.06 \ [95\% \ CI = -0.23–0.35] \pm 0.14)\).

Figure 1 shows the mean \(d'\) values for emotional facial expressions and narratives as a function of the test administered. The figure illustrates the absence of an interaction between the type of task completed by the participants and the test administration, \(F(1, 25) = 0.63, p > .05, \eta_p^2 = .24\). An inspection of the mean \(d'\) values, 95% confidence intervals, and standard errors of the means indicates that participants were not discriminating between the deceptive and truthful emotional facial expressions on the pre-training test \((M = 0.192 \ [95\% \ CI = -0.120–0.503] \pm 0.151)\), but were able to discriminate between them to a small degree on the post-training test \((M = 0.81 \ [95\% \ CI = 0.34–1.28] \pm 0.23)\). This increase in sensitivity for the emotional facial expressions task was found to be significant, \(t(25) = 2.23, p < .05\). Interestingly, for the narratives on the pre-training test, participants were not only failing to discriminate but were labeling the truthful accounts as deceptive, while labeling the deceptive accounts as truthful \((M = -0.40 \ [95\% \ CI = -0.76–(–0.04)] \pm 0.17)\). On the post-training test, however, the participants showed a slight ability to discriminate between the truthful and deceptive narratives \((M = 0.52 \ [95\% \ CI = 0.08–0.97] \pm 0.22)\). This increase in sensitivity was also significant, \(t(25) = 3.36, p < .01\).

Figure 1. Mean values of \(d'\) for the tasks as a function of test. Solid line denotes narratives. Broken line denotes emotional facial expressions. Error bars represent standard error of the mean.
A similar ANOVA was conducted on the $\beta$ scores. The ANOVA did not reveal a main effect of test type on $\beta$ values, $F(1, 25) = 1.8$, $p > .05$, $\eta_p^2 = .07$. The participants' mean values of $\beta$ did not change significantly from the pre-test ($M = -0.08$ [95% CI = $-0.44$–$0.26] \pm 0.17$) to the post-test ($M = 0.27$ [95% CI = $-0.06$–$0.61] \pm 0.16$). The main effect of the type of task on $\beta$ was marginally significant, $F(1, 25) = 3.59$, $p = .07$, $\eta_p^2 = .13$. The participants' obtained lower mean $\beta$ values on the emotional facial expressions task ($M = -0.12$ [95% CI = $-0.41$–$0.17] \pm 0.14$) than on the narratives task ($M = 0.31$ [95% CI = $-0.01$–$0.63] \pm 0.16$). The ANOVA revealed a significant interaction between test and task type, $F(1, 25) = 9.8$, $p < .01$, $\eta_p^2 = .28$ (see Figure 2). Participants exhibited no response bias for facial expressions on either the pre- ($M = 0.06$ [95% CI = $-0.29$–$0.41] \pm 0.17$) or the post-training test ($M = -0.3$ [95% CI = $-0.83$–$0.23] \pm 0.26$), $t(25) = 1.09$, $p > .05$. However, in the case of the narratives, they initially displayed no response bias ($M = -0.23$ [95% CI = $-0.73$–$0.28] \pm 0.24$), but adopted a slightly liberal response bias on the post-training test ($M = 0.85$ [95% CI = $0.35$–$1.35] \pm 0.24$), $t(25) = 2.9$, $p < .01$.

Discussion

This study examined the effectiveness of a brief deception detection training program offered to frontline professionals who assess the credibility of medical claims. The results suggested that the training resulted in modest,
but significant, improvements in their ability to detect deception. First, the training served to improve their overall accuracy rate from chance to above chance on both tasks, with a significant improvement in hit rates specifically. In terms of discrimination ability, the training improved the discriminability roughly equally for emotional expressions and narratives. However, because prior to training sensitivity was negative (below chance) for narratives, after the improvement due to training the $d'$ for narratives was still only modestly positive and less than that for facial expressions.

Although these findings are promising, suggesting that professionals can be rapidly trained to some effect in deception detection, study limitations should be noted when interpreting our results. First, it could be argued that the pre-training test was ‘harder’ than the post-training test in terms of its level of difficulty for the participants, and that this difference was responsible for the observed increases in total accuracy and sensitivity. Although we do not have data to speak directly on this issue, we did make every effort to randomly assign our pool of test stimuli to the pre-training test and post-training test. Second, because of the nature of the training (a within-subjects design), it cannot be concluded with certainty that the apparent training effects were unrelated to practice. Of course, a finding that practice only leads to improved performance would also be important; however, in light of our pattern of results, we doubt that gains in accuracy were due solely to practice or to the possibility that the post-training test was ‘easier’ than the pre-training test. It seems likely that the myth dissolution component of the program was influential in correcting the misperceptions witnessed on the pre-test (for example, by correcting such stereotypes as ‘liars are nervous and truth-tellers are calm’). It should also be noted that an evaluation of the effectiveness of the original training program (on which this brief training was based) dismantled the effects of feedback and information/cue provision via the use of several comparison groups and found that these two training components were related to gains in judgment accuracy over and above practice effects alone (Porter et al., 2000). This finding lends indirect but further support for participation in the brief training program as a valid method of improving deception detection abilities. Nonetheless, future studies including control groups would allow for stronger conclusions and the ability to parse out which components of the program are responsible for the most gains.

Finally, the generalizability of the findings must be considered; because of time constraints, we only studied the discrimination of real and false happiness expressions. Given that other real and false emotions are relevant in high-stakes contexts, they should be studied in the future. For example, the methods utilized by Porter and ten Brinke (2008) to create stimuli of genuine, simulated, neutralized, or masked expressions of disgust, sadness, and fear may be included in future training protocols to assist professionals in differentiating a variety of emotional displays. Further, the use of higher
stakes deception/truthful narratives and facial expressions as stimuli is also needed to examine the generalizability of our findings.

The results from this study, considered with those of previous work suggest that professionals can be trained to improve their deception detection ability. However, the improvements witnessed here are modest and there is much room for improvement. While the brief training approach described here appeared effective in producing modest gains in detecting deception, longer and more comprehensive training sessions (e.g., Porter et al., 2000) appear to produce greater improvements and might be favored where time and resources allow. That said, many professional groups would welcome any increase in deception detection ability. Further practical gains may be achieved with the addition of detailed performance feedback and interactive discussions regarding the integration of verbal, non-verbal and facial cues to deception. Given the consequences of credibility assessment errors in various contexts and our empirical evidence that deception detection can be facilitated, further research is warranted to develop and evaluate empirically based training.

Notes
1. For this analysis, accuracy for the non-offender volunteer injury and offender negative life-event narratives were collapsed into one narratives category.
2. $d'$ is calculated for each participant through the formula, $d' = Z_{\text{Hit Rate}} - Z_{\text{False Alarm Rate}}$. $\beta$ is calculated for each participant through the formula, $\beta = \log(\beta) = -0.5(Z_{\text{Hit Rate}}^2 - Z_{\text{False Alarm Rate}}^2)$.

References


