A test of the Micro-Expressions Training Tool (METT): Does it improve lie detection?

Sarah Jordan[1], Laure Brimbal[2*], D. Brian Wallace[1], Saul M. Kassin[1], Maria Hartwig[1], and Chris N.H. Street[3]

[1] John Jay College of Criminal Justice, City University of New York

[2] Iowa State University

[3] University of Huddersfield

* Corresponding author: W112 Lagomarcino Hall, Ames IA 50010

lbrimbal@iastate.edu

All authors of this manuscript certify that they have no affiliations with or involvement in any organization or entity with any financial interest in the subject matter or materials discussed in this manuscript.

Abstract

The purpose of the study was to examine the effectiveness of the Micro-Expressions Training Tool (METT) in identifying and using micro-expressions to improve lie detection. Participants (n = 90) were randomly assigned to receive training in micro-expressions recognition, a bogus control training, or no training. All participants made veracity judgments of five randomly selected videos of targets providing deceptive or truthful statements. Using Bayesian analyses, we found that the METT group did not outperform those in the bogus training and no training groups. Further, overall accuracy was slightly below chance. Implications of these results are discussed.

Keywords: Micro-expressions, Deception, Lie-detection, Micro-Expressions Training Tool, Training

A test of the Micro-Expressions Training Tool (METT): Does it improve lie detection?

Research on lie detection accuracy has consistently found disappointing results: In a meta-analysis of over 200 studies, Bond and DePaulo (2006) found that people reach on average 54% lie detection accuracy. In another meta-analysis on individual differences in lie detection ability, Bond and DePaulo (2008) found that the variance in accuracy rates does not exceed what would be expected by chance alone (i.e., 50%). Both meta-analyses included studies that tested experts, high and low stakes lies, interaction with the liar, and exposure to liars' baseline behavior. None of these factors had an effect on accuracy. Given that guessing alone produces a 50% accuracy rate, the stable finding of 54% accuracy, combined with little deviation between studies, indicates that human lie detection accuracy is poor (for an examination of the causes of this, see Hartwig & Bond, 2011). Thus, in the current study we sought to answer the important question that is: Is it possible to improve lie detection accuracy with training?

Lie Detection Training

Many different procedures to train deception detection exist; several have been empirically investigated, displaying much variability in their effectiveness (Driskell, 2012; Frank & Feeley, 2003; Meissner & Kassin, 2002). In some cases, training produced substantial improvement (Hartwig, Granhag, Strömwall, & Kronkvist, 2006; Porter, Woodworth, & Birt, 2000) while in others it produced minimal or no increase in accuracy (Köhnken, 1987; Landry & Brigham, 1992; Vrij, 1994). Further, in several studies, training actually reduced accuracy (Kassin & Fong, 1999; Levine, Feeley, McCornack, Huges, & Harms, 2005). A meta-analysis of 16 training studies found that those that included information regarding reliable cues to deception were most effective (Driskell, 2012)—although research has generally failed to find many reliable cues to deception (DePaulo et al., 2003; Hartwig & Bond, 2011).

Interestingly, research shows that training may have an effect independent of the quality of its content. Levine and his colleagues (2005) argued that rather than imparting knowledge relevant to deception detection, training partly affects performance by increasing judges' focus and their critical consideration of the statements. To test this idea, Levine and his colleagues trained participants using valid and bogus training protocols. In line with their argument, they found that bogus and valid training had similar effects on performance. Other research has shown that training increases people's confidence in their judgments as well as the tendency to judge statements as deceptive, when they are otherwise prone to be biased towards seeing the truth (Kassin & Fong, 1999; Masip, Alonso, Garrido, & Herrero, 2009; Meissner & Kassin, 2002). In sum, deception detection training has had a tenuous relationship with improving ability, with a training's effectiveness being influenced not only by the accuracy of its content, but also by the biases and over-confidence it can create in trainees.

Micro-Expressions Training Tool. This study investigates the efficacy of one specific lie detection training program: the Micro-Expressions Training Tool (METT; Ekman, 2006; Paul Ekman Group, 2011). The METT is a self-directed form of training intended to help improve the detection of the micro-expressions sadness, anger, surprise, fear, disgust, contempt, and happiness. Micro-expressions are fleeting facial expressions of felt emotion, which have been reported to last only 1/25 to 1/2 of a second (Ekman, 1985; Matsumoto & Hwang, 2011; Porter, ten Brinke, & Wallace, 2012; Porter & ten Brinke, 2008). The theory behind micro-expressions posits that when people attempt to mask their true emotional state, expressions consistent with their actual state will appear briefly on their face—thus, while people are generally good at hiding their emotions, some facial muscles are more difficult to control than others that is, and automatic displays of emotion will produce briefly detectable emotional "leakage", or micro-

expressions (Ekman, 1985). When a person does not wish to display his or her true feelings s/he will quickly suppress these expressions. Yet there will be an extremely short time between the automatic display of the emotion and the conscious attempt to conceal it, resulting in the micro-expression(s) that can betray a true feeling, and according to theory, aid in detecting deception.

The METT has received some support as a potential tool for improving recognition of overt emotional facial cues. Russell, Chu, and Phillips (1996) found that participants with schizophrenia performed better on an emotional recognition task after receiving the METT, performing to the same degree as control participants not diagnosed with schizophrenia. Marsh et al. (2010) studied longer effects of METT training, and found that emotional recognition was improved a month after training in participants diagnosed with schizophrenia. Matsumoto and Hwang (2011) examined micro-expression training in retail employees. They found that those in the METT group scored better in a test of recognizing micro-expressions and were rated as higher on measures of communicative skill in the work place both immediately after training and two weeks later. In sum, it appears that the METT is able to increase the recognition of overt emotions that people exhibit.

Recognizing micro-expressions may have some utility as an aid for better recognizing facial expressions, but it is more prominently promoted as a potential tool to aid in detecting deception. The METT Advanced program, marketed by the Paul Ekman Group (2011), coined an "online training to increase emotional awareness and detect deception" and promoted with claims that it "…enables you to better spot lies," and "is meant for those whose work requires them to evaluate truthfulness and detect deception—such as police and security personnel" (Paul Ekman Group, METT Advanced-Online only, para. 2). The idea that micro-expression recognition improves lie detection has also been put forth in the scientific literature (Ekman,

2009; Ekman & Matsumoto, 2011; Kassin, Redlich, Alceste, & Luke, 2018) and promoted in the wider culture. One example of this is its use as a focal plot device in the crime drama television series Lie to Me, which ran for three seasons (Baum, 2009). Though a fictional show, Lie to Me was promoted as being based on the research of Ekman. Ekman himself had a blog for the show in which he discussed the science of each episode (Ekman, 2010). Micro-expression recognition training is not only marketed for deception detection, but, more problematically, is actually used for this purpose by the United States government. Training in recognizing micro-expressions is part of the behavioral screening program, known as Screening Passengers by Observation Technique (SPOT) used in airport security (Smith, 2011; Higgenbotham, 2013; Weinberger, 2010). The SPOT program deploys so-called behavior detection officers (BDOs) who receive various training in detecting deception from nonverbal behavior, including training using the METT (the specific content of this program is classified, Higgenbotham, 2013). Evidently, preventing terrorists from entering the country's borders and airports is an important mission. However, to our knowledge, there is no research on the effectiveness of METT in improving lie detection accuracy or security screening efficacy. Given the resources devoted to the SPOT program, evaluating its core component is critical (United States Government Accountability Office [GAO], 2010).

Micro-Expression Training and Lie Detection Accuracy

The research on micro-expressions as a means to detect deception and honesty does not paint an optimistic picture for the utility of METT. Porter and ten Brinke (2008) asked participants to exhibit deceptive or truthful facial expressions in response to emotionally arousing videos. Inconsistent emotional displays did occur in about 50% of the sample, but 22% exhibited partial micro-expressions. These micro-expressions were equally likely to occur in

both feigned and genuine emotional expressions. In a similar study, Porter et al. (2012) also found that true micro-expressions occurred infrequently and were present in both deceptive and honest expressions. They also examined facial expressions in response to high and low emotionally arousing content and found that micro-expressions occur at similar rates in each. These findings were confirmed in an analysis of the facial expressions of high stakes liars and truth tellers (ten Brinke and Porter, 2012).

The presence of micro-expressions also seems to be unrelated to lie detection accuracy. Ekman and O'Sullivan (1991; see also Ekman & Friesen, 1969) found that micro-expression recognition did modestly correlate with lie detection accuracy in a sample of Secret Service agents. Although it is worth noting that the Ekman and O'Sullivan (1991) study has since come under criticism for its selective reporting of the methodology (Bond, 2008). Porter et al. (2012) and Porter and ten Brinke (2008) found untrained observers' ability to detect deception was only marginally better than chance, and was not improved with the presence of micro-expressions in the targets. Warren, Schertler, and Bull (2009) examined the relation between participants' untrained ability to recognize micro-expressions and their ability to detect deception. One set of participants in this study reacted to emotional and neutral video clips. They both fabricated an emotional reaction and reacted genuinely to the videos. A second set of participants viewed these reactions and judged their veracity. After making these judgments, all participants then took the METT post-test in order to measure their accuracy in recognizing micro-expressions. The METT accuracy scores were not correlated with participants' deception judgments. Of note, however, in none of these studies were participants in fact METT trained.

The lack of research on deception detection using micro-expression training is problematic, given the widespread use of METT in practice, in particular in the airport security

screening SPOT program (Higginbotham, 2013; Weinberger, 2010). A report by the Government Accountability Office (GAO) found that there is little empirical validation for the SPOT screening program and that at least 23 individuals attempting terrorist action have passed through screening points without garnering attention (GAO, 2010). Furthermore, this program costs over $200 million a year (GAO, 2010). Hence, it is critical to examine the efficacy of this component of security screening by empirically assessing the effect of METT training on lie detection accuracy.

The Current Study

The purpose of the current study is to test the effect of training in METT on lie detection accuracy. Some participants were randomly assigned to receive METT training; their performance on lie detection tasks was compared to those who received bogus "placebo" training or no training at all. We included a bogus control group in order to account for the effects of lie detection training unrelated to content (i.e., confidence and lie bias; Levine et al., 2005), as discussed above. We also tested METT on deception detection accuracy across different types of truthful and deceptive statements, which were collected from five different research studies. We included multiple types of statements in order to examine the effect of METT training over various types of lies (i.e., high stakes lies (Vrij & Mann, 2001), lies told by convicted felons experience interacting with law enforcement (Kassin et al., 2005; Toomey, 2013), and paradigms that more closely related to airport security screening issues (Street et al., 2011; Sorochinski et al., 2014). High stakes lies were included because some have argued that lie detection accuracy may be higher in situations in which the stakes are high rather than low (Buckley, 2012; Ekman, 1993; O'Sullivan, Frank, Hurley, & Tiwana, 2009). Indeed, someone suspected of an actual crime should experience more emotion and distress than someone who is participating in a

laboratory study. Further, laboratory studies also frequently use students as the targets, who likely do not have any experience lying in high stakes situations. Both of these stipulations could theoretically affect emotional facial displays both true and deceptive. Because the METT is a tool used in airport security screening training, we also wished to utilize statements that were more closely related to issues of airport and border security than the typical theft transgression paradigm (e.g. Hartwig, Granhag, Strömwall, & Vrij, 2005). For instance, airport security officers will be concerned with preventing and apprehending those engaging in terrorist activity and work in border and customs security. We thus included a sample of veracity statements taken from a mock terrorism paradigm (Sorochinski et al., 2014) and a sample of statements made about travel to foreign countries (Street et al., 2011).

Based on the meta-analyses of Bond and DePaulo (2006; 2008), we predict that none of the experimental groups will perform significantly better than chance and again, there will be support for the null hypothesis (Hypothesis 1). Based on previous research on micro-expressions, (Porter et al., 2012; Porter & ten Brinke, 2008; ten Brinke & Porter, 2012; Warren et al., 2009) we likewise predict that there will be evidence supporting the null hypothesis (and not the alternative) when comparing the METT training to bogus and no training conditions on lie detection accuracy (Hypothesis 2). Finally, we predict that training will affect confidence (Hypothesis 3) and lie-bias (Hypothesis 4), whereby participants in the METT and bogus training groups will exhibit higher confidence and a higher lie-bias compared to those who receive no training (Kassin & Fong, 1999; Masip et al., 2009; Meissner & Kassin, 2002).

Method

Participants and Design

Ninety-one students participated in the study. One participant was excluded for failing to follow instructions, yielding a total sample of 90. All participants were students in an introductory psychology course and received course credit for their time. Their mean age was 20.2 (*SD* = 4.3); 71% were female; 29% were male. Participants were randomly assigned to one of the three conditions: METT training (*n* = 30), control training (*n* = 30), or no training (*n* = 30).

Materials

Stimulus videos. Each participant watched one video from each of the five sets of stimuli (five overall), resulting in 450 observations total. Videos were compiled from five different deception detection studies. All sets of videos contained truths and lies; however, videos were randomly selected for each participant resulting in each participant potentially viewing a different number of truths and lies (i.e., some participants saw all truths, some saw all lies, and some saw a combination of truths and lies). Participants were randomly assigned to view the five videos in one of five orders was predetermined by a Latin square. Further, given that certain sets contained more videos than others, we reduced the larger sets by randomly selecting videos from them resulting in a total of 172 videos (thus participants saw certain videos multiple times).

Sorochinski et al. (2014). Videos from this study showed a mock interview of participants trying to convince an interviewer they were not involved in a mock terrorist activity. Suspects in this study had either participated in a mock terrorist activity (liars) or had run an innocuous errand (truth tellers). From this set we obtained 116 statements of which random selection yielded a subsample of 64 videos that we used.

Toomey (2013). Videos from this study showed convicted felons during a mock interview during which they attempted convince the interviewer that they had not stolen a wallet. Participants had either stolen said wallet (liars) or performed a repetitive task (truth tellers).

From this set, we obtained 70 statements; random selection yielded a subsample of 53 videos which we used.

Kassin, Meissner and Norwick (2005). Participants in this study were incarcerated felons who either provided a true or a false confession of their crimes to an interviewer. Participants either confessed to the true reason they were in prison (truth tellers) or were told to confess to a crime they had not committed (liars). From this set we obtained 10 statements, all of which were used in the current study.

Street et al. (2011). The fourth set of videos showed participants who each made two statements describing different countries, one of which they had actually visited (truth) and one which they had never visited (lie). We used all 37 statements obtained from this set.

Vrij and Mann (2001). The fifth set of videos was compiled from press conferences in which individuals made statements asking the public to help find a missing relative or the killer of a relative. For some of these individuals, evidence found after the press conference indicated that they were involved in the disappearance and murder of the person (liars). For others, no such evidence was found and thus they were considered to be telling the truth in their press conference (for a more detailed description of the establishment of ground truth, see Vrij & Mann, 2001). We used all eight statements obtained from this set.

METT module. The METT-trained group took the advanced training module on Ekman's website, "METT Advanced" (Paul Ekman Group, 2011). The METT training consists of five parts: (1) pre-test, (2) training, (3) practice, (4) review, and (5) post-test.

Pre-test. The pre-test presented trainees with 14 human faces (neck-up) in sequence. Each face was shown first with a neutral expression that then flashed an emotional expression lasting 70-100ms (micro-expression) then returned to the original neutral expression. The neutral

face remained on screen until trainees selected which emotion they believed they saw from a list: sadness, anger, surprise, fear, disgust, contempt, or happiness. After classifying all 14 faces, participants were given a score representing how many of the micro-expression emotions were correctly classified.

Training. During the training stage of the METT module, participants were given a list of eight videos, each demonstrating an emotion or combination of emotions: Anger, Disgust, Anger & Disgust I, Anger & Disgust II, Contempt, Happy, Sadness, and Fear & Surprise. Each video displayed a face or pair of faces transitioning between neutral and emotional expression in slow-motion. This was repeated several times while a narrator pointed out specific aspects of the facial expression that were typical of the emotion being displayed. For example, in the Angry video, the narrator noted that, "the eyebrows are pulled down and together in both of these angry expressions".

Practice and review. After the training, participants practiced with 42 videos identical to the pre-test, supplemented with feedback after each selection of emotion displayed. The review consisted of eight videos similar to those in the training phase; the faces in the practice and review videos were different.

Post-test. Finally, trainees took a 28-item post-test. The post-test procedure was identical to the pre-test except that different faces were used. Further, participants were informed that they should try to achieve a score of 80% as this was determined to be the cutoff for proficiency by the training module (Paul Ekman Group, 2011).

Bogus training. We utilized a bogus control training group to account for the non-active effects of training (Levine et al., 2005), such as knowledge of being trained, fatigue effects from time spent to complete the training, and possible effects on motivation and effort. For this

purpose we used the Interpersonal Perception Task (IPT; Costanzo & Archer, 1989) that has been used to train individuals to use verbal and nonverbal behavior to make social perception judgments about other people's inner states and interpersonal relationships (Costanzo, 1992; Costanzo & Archer, 1989). The IPT consists of a 20-minute video in which 15 different scenes are presented. Each scene presents a brief social scenario, in which two or more people were conversing (e.g., an employee and employer talking). One question was posed for each scene, which typically asked the viewer to determine something about the relationship between the people (e.g., identify who in the aforementioned scenario was the employer). The information asked for in the question was not directly presented in the scene, but instead the viewer had to infer the information from the actors' nonverbal behavior. After viewing each scene, the viewer was given a few seconds to answer the related question.

The IPT does address some issues of deception, albeit somewhat briefly (i.e., three scenes) and the majority of the content addresses nonverbal behavior. We do not, however, believe it to be a genuine deception training protocol. It is not put forth as a lie detection training tool and we would not expect it to affect lie detection accuracy, as there is little support for using nonverbal cues to detect deception (DePaulo et al., 2003). However, the idea that nonverbal cues relate to deception is prevalent in the lay public (Akehurst, Köhnken, Vrij, & Bull, 1996) and thus the IPT was well suited as a bogus lie detection training protocol, which is how it was presented to participants.

Procedure

After providing informed consent, participants were randomly assigned to receive the METT module, the bogus training (the IPT), or no training. Those in the METT and IPT conditions were first given instructions informing them that they would be trained in deception

detection. After completing the assigned training (or lack thereof), participants were shown a total of five videos of statements. They were asked to determine if the target in each video was lying or telling the truth and were instructed that for every correct decision they made, they would be entered into a lottery to win $100 –in fact they were all entered in the lottery. Each video was preceded by a short description to provide some contextual information about that specific statement. After each video (five times each), participants completed the dependent measures described below. Once participants had viewed the five videos, they were debriefed, thanked, and dismissed.

Dependent measures. After viewing each stimulus video, participants categorized the target as lying or telling the truth (a dichotomous decision that allowed us to calculate accuracy for each video). Participants then rated their confidence in their decision on a 10-point scale from 1 (*not at all confident*) to 10 (*extremely confident*) and provided a continuous rating of veracity on a 10-point scale from 1 (*definitely telling the truth*) to 10 (*definitely lying*). Finally, participants were asked to provide an open-ended explanation for why they thought the target was lying or telling the truth.

Coding of Reasoning for Veracity Decisions

Two authors independently coded all participants' self-reported reasons for their judgments. The reasons were counted and assigned to one of four a priori categories: Nonverbal-body (NB), nonverbal-face (NF), verbal-content (VC), and verbal-paralingual (VP). For example, "She was moving her hands a lot" was coded as NB; "She kept looking around" was coded as NF; "He gave a lot of detail" was coded as VC; and "He didn't stutter" was coded as VP. Inter-rater reliability between the four categories was adequate, with NB Krippendorff's α = .69, NF Krippendorff's α = .85, VC Krippendorff's α = .78, and VP Krippendorff's α = .69.

Raters' scores were averaged and the means for the four reason categories was calculated across each participant's judgments, and then divided by the length of the response, resulting in four reason-to-word ratios per participant. This was done in order to measure how much explanation and detail was given by participants for their reasons.

Results

METT Training

Of the participants in the METT condition, 37.9% (n = 11) had pre-training scores above the target proficiency score of 80% correct. Post-training proficiency rates were twice as high, with 72.4% (n = 21) of participants above the 80 threshold. The average METT trainee's improvement was significantly larger than zero—an increase of about 12% between pre- and post-training tests, t (28) = 4.11, $p < .001$, M_{diff} = 11.66, 95% CI [5.85, 17.46], dz = 0.76.

Lie Detection Accuracy

Overall, accuracy was slightly but significantly below the chance rate of 50%, (M = .46, 95% CI [.42, .50]), t (89) = -1.98, p = .05, d = 0.21, not supporting our first hypothesis. As an exploratory analysis, we next considered whether lie detection accuracy was uniform among the five samples of statements. Across training conditions, lie detection was poor for statements from all five sets (see Table 1) and were not statistically better than chance. Participants performed significantly worse than chance when judging the statements of Vrij and Mann (2001), exact binomial p (two-tailed) = .031, and Kassin et al. (2005), exact binomial p (two-tailed) = .045. Participants did not perform significantly different from chance when judging the statements of Street et al. (2011), exact binomial p (two-tailed) = .113, Sorochinski et al. (2014), exact binomial p (two-tailed) = .598, and Toomey (2013, exact binomial p (two-tailed) = .598.

We conducted a one-way ANOVA to compare accuracy across experimental conditions (METT vs. IPT vs. no training) to test for differences in lie detection accuracy between groups. This analysis revealed no significant difference, F (2, 87) = 0.22, p = .80, η^2_p = .01, with participants in METT training (M = .46, 95% CI [.38, .55]), IPT training (M = .47, 95% CI [.41, .53]), and no training (M = .44, 95% CI [.37, .50] all performing similarly.

Because a non-significant null hypothesis test does not in fact provide support for the null hypothesis, we calculated Bayes factors to more strictly test our second hypothesis. We used the R package BayesFactor version 0.9.4 (Morey & Rouder, 2013) to calculate a Gunel and Dickey (1974; cited by Jamil et al., 2017) Bayes factor. A Cauchy prior was used with a scaling factor of r = 0.5. The Bayes factor is a ratio of the probability of the null hypothesis against the alternative. Assuming a joint multinomial sampling scheme, the data favored the null hypothesis of no difference between the three conditions by a factor of approximately 21 to 1[1]. That is, we observed strong evidence in favor of there being no difference between conditions.

To determine whether the METT group's performance was attenuated by the nine participants who failed to reach competency on their post-training test we conducted two analyses, both of which led us to believe this was not the case. First, after excluding the nine participants who scored under 80, METT-trained participants' performance increased slightly to just under the chance rate (M = .49, 95% CI [.40, .59]), but there remained no significant differences between conditions in lie-detection accuracy when submitted to the same one-way ANOVA as described above, F (2, 81) = 0.47, p = .63, η^2_p = .012. The Bayes factor also

[1] A ratio of the probabilities of two hypotheses closer to 1 to 1 (or Bayes Factor of 1) indicates a lack of evidence, while those that diverge from 1 to 1 indicate evidence in favor of one of the two hypotheses (alternative and null), here with increasing odds (or higher value Bayes Factors) indicating higher likelihood of the data occurring under the null hypothesis (for a more detailed explanation of Bayes Factors, see Jarosz & Wiley, 2014).

indicated a higher probability of the null versus the alternative hypothesis being true by a factor of 17 to 1. Second, lie-detection accuracy within the METT condition was not significantly correlated with the METT proficiency pre-test score, r (27) = .16, p = .41, 95% CI [-.14, .41] or the post-test final score, r (27) = .24, p = .21, 95% CI [-.21, .64]. A Bayes factor gave indeterminate results and could not differentiate whether the null of no correlation was more probable than the alternative hypothesis of there being a correlation, given the observed data (Bayes factors of 2.09 and 1.16, respectively; see Ly, Verhagen & Wagenmakers, 2016). A METT performance change score was calculated, to reflect to what extent METT-condition participants improved on the METT proficiency test administered before and after training. Lie-detection accuracy was not significantly correlated with this METT change score, r (27) = -.02, p = .91, 95% CI [-.28, .23]. A Bayes Factor of 4 favored the null hypothesis of no correlation.

Across all five samples of statements, participants who received METT training did not perform differently from those who received bogus or no training (see Table 1). All three groups performed equally poorly when judging the real-world, high stakes statements of Vrij and Mann (2001), χ^2 (2, N = 87) = 1.01, p = .610, φ_c = .11. The groups did not significantly differ in their judgments of airport screening-related statements from Sorochinski et al. (2014), χ^2(2, N = 90) = 0.27, p = .875, φ_c = .05, or Street et al. (2011), χ^2(2, N = 90) = 0.64, p = .725, φ_c = .08. All of the groups performed equally poorly judging statements made by participants with felony records from Toomey (2013), χ^2 (2, N = 90) = 0.27, p = .875, φ_c = .05, or from Kassin et al. (2005), χ^2 (2, N = 90) = 2.90, p = .235, φ_c = .18.

Confidence in Judgment

Participants were highly confident overall (M = 7.25 on the 10-point scale, 95% CI [7.00, 7.51]). A 3 (training: METT, IPT, none) x 2 (statement veracity: truth or lie) mixed ANOVA

found no main effect of training, F (2, 84) = 0.82, p = .44, η^2_p = .019 –all groups were highly confident. Though no more accurate in their classifications, participants were significantly, but only slightly, more confident after evaluating a truthful statement (M = 7.35, 95% CI [6.87, 7.83]) than after judging a deceptive one (M = 6.93, 95% CI [6.65, 7.40]), F (1, 84) = 4.93, p = .03, d = 0.23. The training by statement veracity interaction was not significant, F (2, 84) = 0.87, p = .42, η^2_p = .020. The confidence-accuracy relationship was explored through a linear regression, with moderators to determine if the confidence-accuracy relationship differed by experimental condition. The model included lie-detection accuracy as the outcome and the following predictor variables (see Table 1): decision confidence, experimental condition (dummy coded with the control as reference), and the condition by confidence interaction. The confidence-accuracy relationship in the control condition did not significantly differ from zero, β = -.039, t (84) = -1.24, p = .219, and the moderator terms indicated that this confidence-accuracy relationship did not significantly differ between the control and IPT, β = .058, t (84) = -1.40, p = .165 nor the control and METT, β = .007, t (84) = -.145, p = .885.

Continuous Veracity Measure

A 3 (training: METT, IPT, or none) x 2 (statement veracity: truth or lie) mixed ANOVA was performed to investigate the extent to which participants' continuous judgments of statement truthfulness differed by condition and by the veracity of the statement. Higher means indicated more perceived deception. There was no significant omnibus main effect of training, F (2, 84) = 2.78, p = .07, η^2_p = .062. A Bayes factor was calculated to compare two logistic mixed effects models. The alternative model replicated the preceding ANOVA, while the null model removed the main effect of training. A scaling factor of r = 1.0 was used to fit the Cauchy prior to the nuisance variables of participant number and video number. The Bayes factor favored the

null hypothesis of no effect of training by a factor of 4.

We next compared METT and IPT trained participants' judgment biases to those displayed by untrained participants. Lie bias was similar in the three conditions as well, F (2, 87) = 0.58, p = .56, η^2_p = .013, with no significant differences found in the number of lie judgments between METT (M = 2.47, 95% CI [2.07, 2.87]), IPT (M = 2.20, 95% CI [1.77, 2.63]), and no training (M = 2.20, 95% CI [1.79, 2.61]). A Bayes factor on the contingency table shifted the relative plausibility of the null compared to the alternative by a factor of 14.

Reasoning for Veracity Decisions

As an exploratory analysis, we examined participants reasoning for their decisions. We conducted a 3 (training: METT, IPT, none) x 4 (reason: NB, NF, VC, VP) mixed ANOVA, with the reason-to-word ratio as the dependent variable. Sphericity was violated (Mauchly's W = .846, p = .15), thus Greenhouse-Geisser corrections to the F ratios are reported below. There was no significant main effect of the training, F (2, 86) = 1.79, p = .174, η^2_p = .040, as all three groups reported similar reason-to-word ratios overall. There was no significant main effect of reasoning, F (2.7, 233.3) = 1.72, p = .168, η^2_p = .020 –each of the four reason categories were observed at similar rates overall. Although the training by reason interaction was also not significant, F (5.4, 233.3) = 1.24, p = .289, η^2_p = .028, we performed a planned contrast of the differences between training conditions in the non-verbal facial category. Of the four reason categories, we expected METT participants to include a larger number of non-verbal facial reasons for their veracity judgments. The results support our hypothesis. Moderate effect sizes were observed indicating that the METT-trained group (M = 3.31, SD = 2.06) cited more facial reasons than IPT-trained participants (M = 2.37, SD = 2.07), p = .064, d_z = 0.46, and untrained participants (M = 2.16, SD = 1.67), p = .025, d_z = 0.61.

Discussion

This study is the first known test of the METT as a lie detection tool and offers no empirical support for its effectiveness. This conclusion is based on two sources of evidence. First, METT-trained individuals performed worse than chance; guessing would have produced marginally better results. Second, METT-trained participants performed no better than untrained or bogus trained individuals, in fact, there was strong support for the null hypothesis. All groups were very highly and equally confident in their ability to detect deception, even though their performance was poor overall. Prior research has demonstrated that training can lead to a lie bias, or a propensity to judge veracity statements as lies more frequently (Blair, 2006; Masip et al., 2009; Meissner & Kassin, 2004). We did not find support for this response bias in dichotomous judgments of truth and deception. We did, however, find evidence for it when participants were asked to rate their impressions of truthfulness on a continuous measure. This may be due to the fact that the impression ratings allowed for more range, and thus were more sensitive to capturing the effect than the dichotomous judgments. Further, when asked to rate the statements' truthfulness, the METT group rated statements as less truthful overall than the untrained control group—with a moderate effect size (.40). So, while their training failed to improve their lie-detection abilities, it did make them more inclined to think people were lying.

In anticipation of the possibility that the effectiveness of METT is task-specific, we also examined the effect of training across five different types of veracity statements. Our sets of videos provided for a range of stimuli in terms of content (i.e., relevant to security screening, missing persons, past transgressions), speakers (i.e., college students, prison inmates, people with a criminal record), and stakes (i.e., laboratory and field settings, real-world conditions). Yet, the

METT was no more effective than bogus or no training when assessing different types of lies, with accuracy still no better than chance for judging each type of statement.

Despite the claims made and popular conceptions, the failure of the METT to bolster lie-detection performance is not unexpected. Prior research shows that to the extent that micro-expressions even occur, they are rare and occur both in the presence of truth telling and lying (Porter et al., 2012; Porter & ten Brinke, 2008). Thus, even if the METT is successful in improving the detection of emotions in micro-expressions (a proposition we did not test), there is no a priori reason to believe that these emotions are valid indicators of truthfulness or deception. Training people in explicitly recognizing micro-expressions (via the METT) did not significantly improve lie detection accuracy (using Null-Hypothesis Significance Testing; NHST) and Bayesian analyses provided strong support for the lack of an effect, further supporting the proposition that micro-expressions are not linked to deception. The findings of this study are also in keeping with the vast majority of previous lie detection research, showing that people are little better than chance when making veracity judgments (Bond & DePaulo, 2006). Further, we build upon previous research with the use of Bayesian analyses providing the beneficial ability to test, and indeed find strong support for the lack of effectiveness of METT to detect deception, a proposition that cannot be supported using more standard NHST (for more detailed accounts of the limitations of NHST, see Rouder, Speckman, Sun, Morey, & Iverson, 2009; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011).

Limitations

There are some limitations to our study. Our study involved judgments made by college students from videos played in a laboratory setting. However, prior research has shown that observers with access to a live target do not perform significantly better (Hartwig, Granhag,

Strömwall, & Vrij, 2004); nor do police and other professionals perform significantly better than students (Bond & DePaulo, 2006; DePaulo & Pfeifer, 1986; Vrij, Akehurst, Brown, & Mann, 2006). Further, even if these factors limit the effectiveness of the METT, we should still expect improvement over no training or bogus training. Of the participants in our METT training group 30% did not meet the 80% proficiency score while nearly 38% of them met it on the pre-test measure. Taken together, it would seem that the METT did not make meaningful changes in the participants and would thus explain our findings. However, there is some evidence that METT training did have an impact on how participants approached the task. For one, participants who received METT training were more likely to focus on facial cues relative to other groups. Second, METT trained participants exhibited an improved ability to recognize micro-expressions from the pre-test to the post-test. However, it could be argued that improvement observed in recognizing micro-expressions (as expressed by the pre-test to post- test score improvements) could be due to practice effects rather than an actual improvement in skill. We see these as limitations as a function of the METT's design rather than our research paradigm, as our participants took the test in the same program directed manner that consumers who purchase the test would. In fact this further demonstrates the ineffectiveness of the design of the METT. When we removed those who did not meet proficiency from the analysis, accuracy did not improve for the METT group and accuracy was not correlated with proficiency score. So even those who tested as highly skilled in recognizing micro-expressions did not perform better than chance in detecting deception.

Conclusions

The Paul Ekman Group website claims that the METT "enables you to better able to spot lies" (Paul Ekman Group, What are Micro Expressions?, para. 1) and "has been scientifically proven and field tested" (Paul Ekman Group, How to Get Started, para. 4). We are not aware of this scientific proof for lie detection efficacy, so we aimed to directly test the claim. Our findings do not support the use of METT as a lie detection tool. The METT did not improve accuracy any more than a bogus training protocol or even no training at all. The METT also did not improve accuracy beyond the level associated with guessing. This is problematic to say the least given that training in the recognition of micro-expressions comprises a large part of a screening system that has become ever more pervasive in our aviation security (Higginbotham, 2013; Weinberger, 2010).

References

Akehurst, L., Köhnken, G., Vrij, A., & Bull, R. (1996). Lay persons' and police officers' beliefs regarding deceptive behaviour. *Applied Cognitive Psychology*, *10*, 461-471.

Baum, S. (Producer). (2009). *Lie to me*. Hollywood: Imagine Entertainment 20th Century Fox Television

Blair, J. P. (2006). From the field: Can detection of deception response bias be manipulated? *Journal of Crime and Justice*, *29*, 141-152.

Bond, C. F. (2008). A few can catch a liar, sometimes: Comments on Ekman and O'Sullivan (1991), as well as Ekman, O'Sullivan, and Frank (1999). *Applied Cognitive Psychology*, *22*, 1298-1300.

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

Bond, C. F. & DePaulo, B. M. (2008). Individual differences in judging deception: Accuracy and bias. *Psychological Bulletin*, *134*, 477-492.

Buckley, J. P. (2012). Detection of deception researchers needs to collaborate with experienced practitioners. *Journal of Applied Research in Memory and Cognition*, *1*, 126-127.

Costanzo, M. (1992). Training students to decode verbal and nonverbal cues: Effects on confidence and performance. *Journal of Educational Psychology*, *84*, 308-313.

Costanzo, M. & Archer, D. (1989). Interpreting the expressive behavior of others: The interpersonal perception task. *Journal of Nonverbal Behavior*, *13*, 225-245.

Ekman, P. (1992). *Telling lies: Clues to deceit in the marketplace, politics, and marriage*. New York: Norton and Company.

DePaulo, B. M., Lindsay, J. L., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin, 129,* 74-118.

DePaulo, B. M., & Pfeifer, R. L. (1986). On-the-job experience and skill at detecting deception. *Journal of Applied Social Psychology, 16,* 249-267.

Driskell, J. E. (2012). Effectiveness of deception detection training: A meta-analysis. *Psychology, Crime & Law, 18,* 713-731.

Ekman, P. (1985). *Telling lies: Clues to deceit in the marketplace, marriage, and politics*. New York: Norton.

Ekman, P. (1993). Why don't we catch liars? *Social Research, 63,* 801-807.

Ekman, P. (2006, October 29). How to spot a terrorist on the fly. The Washington Post. Retrieved from http://www.washingtonpost.com/wp-dyn/content/article/2006/10/27/AR2006102701478.html

Ekman, P. (2009). Lie catching and microexpressions. In C. Martin (Ed.), *The Philosophy of Deception* (pp. 118-135). Oxford: Oxford University Press.

Ekman, P. (2010). Lie to me blog. [Web log]. Retrieved from http://www.lie-to-me-now.com/blog

Ekman, P., & Friesen, W. V. (1969). Nonverbal leakage and clues to deception. *Psychiatry, 32,* 88-105.

Ekman, P., & Matsumoto, D. (2011). Reading faces: The universality of emotional expression. In M. Gernsbacher, R. W. Pew, L. M. Hough, J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 140-146). New York, NY: Worth Publishers.

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

Frank, M. G. & Feeley, T. H. (2003). To catch a liar: Challenges for research in lie detection training. *Journal of Applied Communication Research, 31,* 58-75.

Government Accountability Office. (2010). Efforts to validate TSA's passenger screening behavior detection program underway, but opportunities exist to strengthen validation and address operational challenges (GAO-10-763). Retrieved from http://www.gao.gov/products/GAO-10-763

Hartwig, M., & Bond, C. F., Jr. (2011). Why do lie-catchers fail? A lens model meta-analysis of human lie judgments. *Psychological Bulletin, 137,* 643-659.

Hartwig, M., Granhag, P., Strömwall, L. A., & Kronkvist, O. (2006). Strategic use of evidence during police interviews: When training to detect deception works. *Law and Human Behavior, 30,* 603-619.

Hartwig, M., Granhag, P. A., Strömwall, L. A., & Vrij, A. (2004). Police officers' lie detection accuracy: Interrogating freely versus observing video. *Police Quarterly, 7,* 429-456.

Hartwig, M., Granhag, P.A., Strömwall, L.A., & Vrij, A. (2005). Detecting deception via the strategic disclosure of evidence. *Law and Human Behavior, 29,* 469-484.

Higginbotham, A. (2013). Deception is futile when big brother's lie detector turns its eyes on you. *Wired, 21,* 90-97.

Jamil, T., Ly, A., Morey, R. D., Love, J., Marsman, M., & Wagenmakers, E.-J. (2017). Default "Gunel and Dickey" Bayes factors for contingency tables. *Behavior Research Methods, 49,* 638-652.

Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes Factors. *Journal of Problem Solving, 7,* 2-9.

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.

Kassin, S. M., Meissner, C. A., & Norwick, R. J. (2005). 'I'd know a false confession if I saw one': A comparative study of college students and police investigators. *Law and Human Behavior, 29,* 211-227.

Kassin, S. M., Redlich, A. D., Alceste, F., & Luke, T. J. (2018). On the general acceptance of confessions research: Opinions of the scientific community. *American Psychologist, 73,* 63-80.

Köhnken, G. (1987). Training police officers to detect deceptive eyewitness statements: Does it work?. *Social Behaviour, 2,* 1-17.

Landry, K. L. & Brigham, J. C. (1992). The effect of training in criteria-based content analysis on the ability to detect deception in adults. *Law and Human Behavior, 16,* 663-676.

Levine, T. R., Feeley, T., McCornack, S. A., Hughes, M., & Harms, C. M. (2005). Testing the effects of nonverbal behavior training on accuracy in deception detection with the inclusion of a bogus training control group. *Western Journal of Communication, 69,* 203-217.

Ly, A., Verhagen, J., & Wagenmakers, E.-J. (2016). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology, 72,* 19-32.

Marsh, P. J., Green, M. J., Russell, T. A., McGuire, J., Harris, A., Coltheart, M. (2010). Remediation of facial emotion recognition in schizophrenia: Functional predictors, generalizability, and durability. *American Journal of Psychiatric Rehabilitation, 13,* 143-170.

Masip, J., Alonso, H., Garrido, E., & Herrero, C. (2009). Training to detect what? The biasing effect of training on veracity judgments. *Applied Cognitive Psychology, 23,* 1282-1296.

Matsumoto, D. & Hwang, H. S. (2011). Evidence for training the ability to read microexpressions. *Motivation and Emotion, 35,* 181-191.

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

Morey, R. D., & Rouder, J. (2013). Bayesfactor: *An R package for Bayesian analysis in common research design*. Retrieved 15 January 2015 from http://bayesfactorpcl.r-forge.r-project.org.

O'Sullivan, M., Frank, M. G., Hurley, C. M., & Tiwana, J. (2009). Police lie detection accuracy: The effect of lie scenario. *Law and Human Behavior, 33,* 530-538.

Paul Ekman Group (2011). F.A.C.E. training: Interactive training by Dr. Paul Ekman. Retrieved from https://face.paulekman.com

Porter, S. & ten Brinke, L. (2008). Reading between the lies: Identifying concealed and falsified emotions in universal facial expressions. *Psychological Science, 19,* 508-514.

Porter, S., ten Brinke, L., & Wallace, B. (2012). Secrets and lies: Involuntary leakage in deceptive facial expressions as a function of emotional intensity. *Journal of Nonverbal Behavior, 36,* 23-37.

Porter, S., Woodworth, M., & Birt, A. R. (2000). Truth, lies, and videotape: An investigation of the ability of federal parole officers to detect deception. *Law and Human Behavior, 24,* 643-658.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 56,* 356-374.

Russell, T. A., Chu, E., & Phillips, M. L. (2006). A pilot study to investigate the effectiveness of emotion recognition remediation in schizophrenia using the micro-expression training tool. *British Journal of Clinical Psychology*, *45*, 579-583.

Smith, T. (2011, August 16). Next in line for the TSA? A thorough 'chat-down'. *National Public Radio*. Retrieved from http://www.npr.org/2011/08/16/139643652/next-in-line-for-the-tsa-a-thorough-chat-down

Sorochinski, M., Hartwig, M., Osborne, J., Wilkins, E., Marsh, J. J., Kazakov, D., & Granhag, P. A. (2014). Interviewing to detect deception: When to disclose the evidence? *Journal of Police and Criminal Psychology*, *29*, 87-94.

Street, C., Tbaily, L. Baron, S., Khalil-Marzouk, P., Wright, K., Hanby, B., & Richardson, D. C. (2011). *Bloomsbury Deception Set*. Paper presented at the BPS Division of Forensic Psychology Conference, Portsmouth, UK.

Ten Brinke, L. & Porter, S. (2012). Cry me a river: Identifying the behavioral consequences of extremely high-stakes interpersonal deception. *Law and Human Behavior*, *36,* 469-477.

Toomey, J. A. (2013). *Investigating the role of psychopathic characteristics in deceptive behavior* (Unpublished doctoral dissertation). The Graduate Center, City University of New York, New York, NY.

Vrij, A. (1994). The impact of information and setting on deception detection by police detectives. *Journal of Nonverbal Behavior*, *18*, 117-136.

Vrij, A., Akehurst, L., Brown, L., & Mann, S. (2006). Detecting lies in young children, adolescents, and adults. *Applied Cognitive Psychology, 20,* 1225-1237.

Vrij, A. & Mann, S. (2001). Who killed my relative? Police officers' ability to detect real-life high-stakes lies. *Psychology, Crime & Law, 7,* 119-132.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. (2011). Why psychologists must change the way they analyse their data: The case of Psi: Comment on Bem (2011). *Journal of Personality and Social Psychology, 100,* 426-432.

Warren, G., Schertler, E., & Bull, P. (2009). Detecting deception from emotional and unemotional cues. *Journal of Nonverbal Behavior, 33,* 59-69.

Weinberger, S. (2010). Airport security: Intent to deceive?. *Nature, 465,* 412-415.

Table 1

Lie detection percent accuracy by condition and video set

	Training			
Video set	METT	IPT	No training	Total
Vrij & Mann (2001)	44.80	32.10	36.70	37.90
Street et al. (2011)	53.50	63.30	60.00	58.90
Sorochinski et al. (2014)	43.30	50.00	46.70	46.70
Toomey (2013)	46.70	43.30	50.00	46.70
Kassin et al. (2005)	43.30	46.70	26.70	38.90
Total	46.30	47.30	44.00	45.90